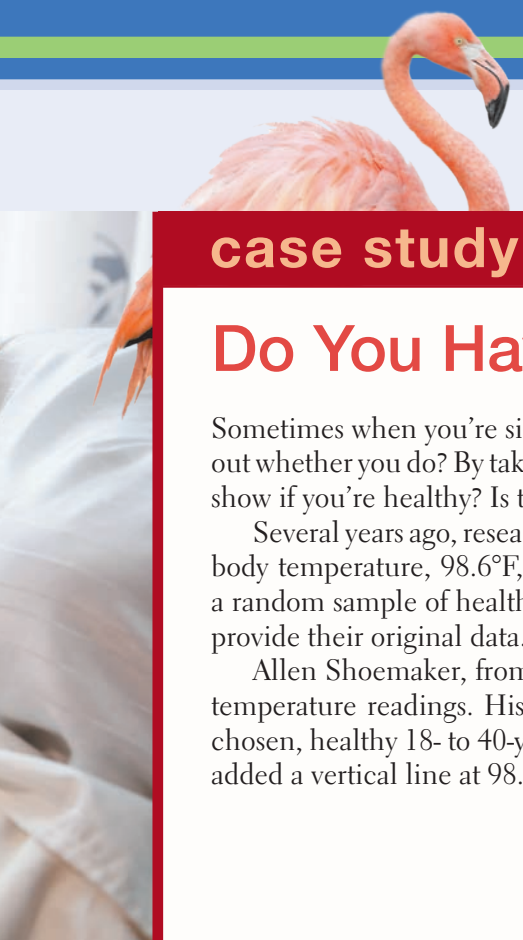
Testing a Claim



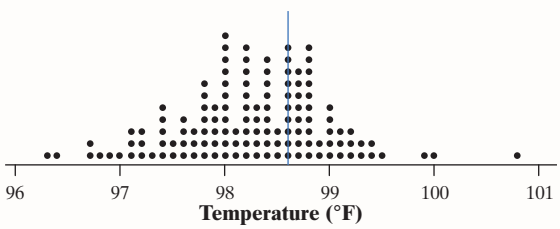
case study

Do You Have a Fever?

Sometimes when you're sick, your forehead feels really warm. You might have a fever. How can you find out whether you do? By taking your temperature, of course. But what temperature should the thermometer show if you're healthy? Is this temperature the same for everyone?

Several years ago, researchers conducted a study to determine whether the "accepted" value for normal body temperature, 98.6°F, is accurate. They used an oral thermometer to measure the temperatures of a random sample of healthy men and women aged 18 to 40. As is often the case, the researchers did not provide their original data.

Allen Shoemaker, from Calvin College, produced a data set with the same properties as the original temperature readings. His data set consists of one oral temperature reading for each of 130 randomly chosen, healthy 18- to 40-year-olds.[1] A dotplot of Shoemaker's temperature data is shown below. We have added a vertical line at 98.6°F for reference.



Exploratory data analysis revealed several interesting facts about this data set:

- The mean temperature was $\bar{x} = 98.25$°F.
- The standard deviation of the temperature readings was $s_x = 0.73$°F.
- 62.3% of the temperature readings were less than 98.6°F.

Based on the results of this study, can we conclude that "normal" body temperature in the population of healthy 18- to 40-year-olds is *not* 98.6°F? By the end of this chapter, you will have developed the necessary tools for answering this question.

Introduction

Confidence intervals are one of the two most common types of statistical inference. Use a confidence interval when your goal is to estimate a population parameter. The second common type of inference, called *significance tests*, has a different goal: to assess the evidence provided by data about some claim concerning a parameter. Here is an Activity that illustrates the reasoning of statistical tests.

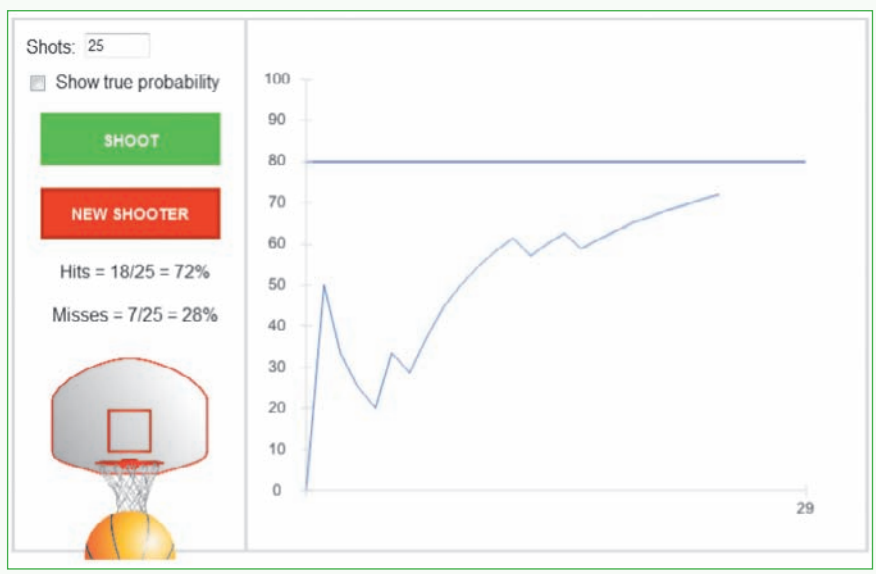
ACTIVITY I'm a Great Free-Throw Shooter!

MATERIALS:

Computer with Internet access and projection capability

A basketball player claims to make 80% of the free throws that he attempts. We think he might be exaggerating. To test this claim, we'll ask him to shoot some free throws—virtually—using *The Reasoning of a Statistical Test* applet at the book's Web site.

1. Go to www.whfreeman.com/tps5e and launch the applet.



2. Set the applet to take 25 shots. Click "Shoot." How many of the 25 shots did the player make? Do you have enough data to decide whether the player's claim is valid?
3. Click "Shoot" again for 25 more shots. Keep doing this until you are convinced *either* that the player makes less than 80% of his shots *or* that the player's claim is true. How large a sample of shots did you need to make your decision?
4. Click "Show true probability" to reveal the truth. Was your conclusion correct?
5. If time permits, choose a new shooter and repeat Steps 2 through 4. Is it easier to tell that the player is exaggerating when his actual proportion of free throws made is closer to 0.8 or farther from 0.8?

In the free-throw shooter Activity, the parameter of interest is the proportion p of free throws that the player will make if he shoots forever. Our player claims that $p = 0.80$. To test his claim, we let the applet simulate 25 shots. If the player makes only 40% of his free throws (10 of 25 shots made), we have fairly strong evidence that he doesn't shoot 80%. But what if he makes 76% of his free throws (19 of 25 shots made)? This provides *some* evidence that his true long-term percent may be less than 80%, but it's not nearly as convincing as $\hat{p} = 0.40$. Statistical tests weigh the evidence *against* a claim like $p = 0.8$ and in favor of a counter-claim like $p < 0.80$.

Section 9.1 focuses on the underlying logic of statistical tests. Once the foundation is laid, we consider the implications of using these tests to make decisions—about everything from free-throw shooting to the effectiveness of a new drug. In Section 9.2, we present the details of performing a test about a population proportion. Section 9.3 shows how to test a claim about a population mean. Along the way, we examine the connection between confidence intervals and tests.

9.1 Significance Tests: The Basics

WHAT YOU WILL LEARN By the end of the section, you should be able to:

- State the null and alternative hypotheses for a significance test about a population parameter.
- Interpret a *P*-value in context.
- Determine whether the results of a study are statistically significant and make an appropriate conclusion using a significance level.
- Interpret a Type I and a Type II error in context and give a consequence of each.

A **significance test** is a formal procedure for using observed data to decide between two competing claims (also called *hypotheses*). The claims are often statements about a parameter, like the population proportion p or the population mean μ. Let's start by taking a closer look at how to state hypotheses.

Stating Hypotheses

A significance test starts with a careful statement of the claims we want to compare. In our free-throw shooter example, the virtual player claims that his long-run proportion of made free throws is $p = 0.80$. This is the claim we seek evidence *against*. We call it the **null hypothesis**, abbreviated H_0. Usually, the null hypothesis is a statement of "no difference." For the free-throw shooter, no difference from what he claimed gives $H_0: p = 0.80$.

Remember: the null hypothesis is the dull hypothesis!

The claim we hope or suspect to be true instead of the null hypothesis is called the **alternative hypothesis**. We abbreviate the alternative hypothesis as H_a. In this case, we believe the player might be exaggerating, so our alternative hypothesis is $H_a: p < 0.80$.

DEFINITION: Null hypothesis H_0, alternative hypothesis H_a

The claim we weigh evidence against in a statistical test is called the **null hypothesis (H_0)**. Often the null hypothesis is a statement of "no difference."

The claim about the population that we are trying to find evidence *for* is the **alternative hypothesis (H_a)**.

Some people insist that all three possibilities—greater than, less than, and equal to—be accounted for in the hypotheses. For the free-throw shooter example, since the alternative hypothesis is $H_a: p < 0.80$, they would write the null hypothesis as $H_0: p \geq 0.80$. In spite of the mathematical appeal of covering all three cases, we use only the value $p = 0.80$ in our calculations. So we'll stick with $H_0: p = 0.80$.

In the free-throw shooter example, our hypotheses are

$$H_0: p = 0.80$$
$$H_a: p < 0.80$$

where p is the long-run proportion of made free throws. The alternative hypothesis is **one-sided** because we are interested only in whether the player is overstating his free-throw shooting ability. Because H_a expresses the effect that we hope to find evidence *for*, it is sometimes easier to begin by stating H_a and then set up H_0 as the statement that the hoped-for effect is not present.

Here is an example in which the alternative hypothesis is **two-sided**.

EXAMPLE Juicy Pineapples

Stating hypotheses

At the Hawaii Pineapple Company, managers are interested in the size of the pineapples grown in the company's fields. Last year, the mean weight of the pineapples harvested from one large field was 31 ounces. A different irrigation system was installed in this field after the growing season. Managers wonder how this change will affect the mean weight of pineapples grown in the field this year.

PROBLEM: State appropriate hypotheses for performing a significance test. Be sure to define the parameter of interest.

SOLUTION: The parameter of interest is the mean weight μ of all pineapples grown in the field this year. Because managers wonder whether the mean weight of this year's pineapples will differ from last year's mean weight of 31 ounces, the alternative hypothesis is two-sided; that is, either $\mu < 31$ or $\mu > 31$. For simplicity, we write this as $\mu \neq 31$. The null hypothesis says that there is no difference in the mean weight of the pineapples after the irrigation system was changed. That is,

$$H_0: \mu = 31$$
$$H_a: \mu \neq 31$$

For Practice *Try Exercise* 1

The hypotheses should express the hopes or suspicions we have ***before*** *we see the data. It is cheating to look at the data first and then frame hypotheses to fit what the data show.* For example, the data for the pineapple study

showed that $\bar{x} = 31.935$ ounces for a random sample of 50 pineapples grown in the field this year. You should *not* change the alternative hypothesis to $H_a: \mu > 31$ after looking at the data. If you do not have a specific direction firmly in mind in advance, use a two-sided alternative hypothesis.

DEFINITION: One-sided alternative hypothesis and two-sided alternative hypothesis

The alternative hypothesis is **one-sided** if it states that a parameter is *larger than* the null hypothesis value or if it states that the parameter is *smaller than* the null value. It is **two-sided** if it states that the parameter is *different from* the null hypothesis value (it could be either larger or smaller).

It is common to refer to a significance test with a one-sided alternative hypothesis as a *one-sided test* or *one-tailed test* and to a test with a two-sided alternative hypothesis as a *two-sided test* or *two-tailed test*.

The null hypothesis has the form H_0: parameter = value. The alternative hypothesis has one of the forms H_a: parameter < value, H_a: parameter > value, or H_a: parameter ≠ value. To determine the correct form of H_a, read the problem carefully.

Hypotheses always refer to a *population*, not to a sample. Be sure to state H_0 and H_a in terms of *population parameters*. It is *never* correct to write a hypothesis about a sample statistic, such as $\hat{p} = 0.64$ or $\bar{x} > 85$.

CHECK YOUR UNDERSTANDING

For each of the following settings, (a) describe the parameter of interest, and (b) state appropriate hypotheses for a significance test.

1. According to the Web site sleepdeprivation.com, 85% of teens are getting less than eight hours of sleep a night. Jannie wonders whether this result holds in her large high school. She asks an SRS of 100 students at the school how much sleep they get on a typical night. In all, 75 of the responders said less than 8 hours.

2. As part of its 2010 census marketing campaign, the U.S. Census Bureau advertised "10 questions, 10 minutes—that's all it takes." On the census form itself, we read, "The U.S. Census Bureau estimates that, for the average household, this form will take about 10 minutes to complete, including the time for reviewing the instructions and answers." We suspect that the actual time it takes to complete the form may be longer than advertised.

The Reasoning of Significance Tests

Significance tests ask if sample data give convincing evidence against the null hypothesis and in favor of the alternative hypothesis. A test answers the question, "How likely is it to get a result like this just by chance when the null hypothesis is true?" The answer comes in the form of a probability.

A significance test is sometimes referred to as a *test of significance*, a *hypothesis test*, or a *test of hypotheses*.

Here is an activity that introduces the underlying logic of significance tests.

ACTIVITY | I'm a Great Free-Throw Shooter!

MATERIALS:
Copy of pie chart with 80% shaded and paper clip for each student

Our virtual basketball player in the previous Activity claimed to be an 80% free-throw shooter. Suppose that he shoots 50 free throws and makes 32 of them. His sample proportion of made shots is $\hat{p} = \frac{32}{50} = 0.64$. This result suggests that he may really make less than 80% of his free throws in the long run. But do we have convincing evidence that $p < 0.80$? In this activity, you and your classmates will perform a simulation to find out.

1. Make a spinner that gives the shooter an 80% chance of making a free throw. Using the pie chart provided by your teacher, label the 80% region "made shot" and the 20% region "missed shot." Straighten out one of the ends of a paper clip so that there is a loop on one side and a pointer on the other. On a flat surface, place a pencil through the loop, and put the tip of the pencil on the center of the pie chart. Then flick the paper clip and see where the pointed end lands.
2. Simulate a random sample of 50 shots. Flick the paper clip 50 times, and count the number of times that the pointed end lands in the "made shot" region.
3. Compute the sample proportion $\hat{p}$ of made shots in your simulation from Step 2. Plot this value on the class dotplot drawn by your teacher.
4. Repeat Steps 2 and 3 as needed to get at least 40 trials of the simulation for your class.
5. Based on the results of your simulation, how likely is it for an 80% shooter to make 64% or less when he shoots 50 free throws? Does the observed $\hat{p} = 0.64$ result give convincing evidence that the player is exaggerating?

Figure 9.1 shows what sample proportions are likely to occur by chance alone, assuming that $p = 0.80$.

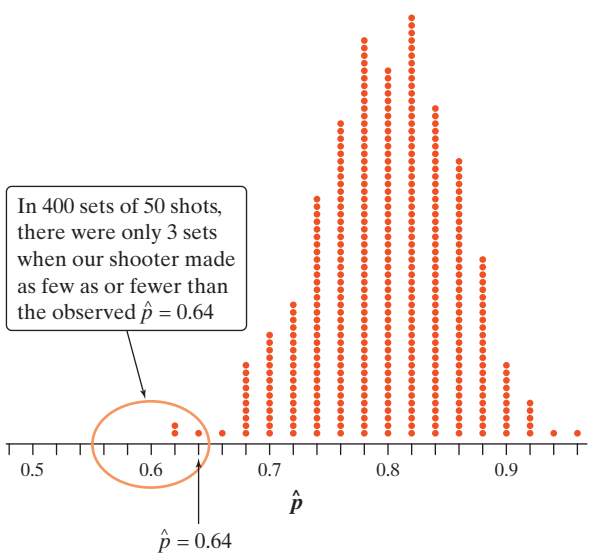


FIGURE 9.1 Fathom dotplot of the sample proportion $\hat{p}$ of free throws made by an 80% shooter in 400 sets of 50 shots.

Our reasoning in the Activity is based on asking what would happen if the player's claim ($p = 0.80$) were true and we observed many samples of 50 free throws. We used Fathom software to simulate 400 sets of 50 shots assuming that the player is really an 80% shooter. Figure 9.1 shows a dotplot of the results. Each dot on the graph represents the proportion of made shots in one set of 50 attempts. For example, if the player makes 43/50 shots in one trial, the dot would be placed at $\hat{p} = 0.86$.

You can say how strong the evidence against the player's claim is by giving the probability that he would make as few as 32 out of 50 free throws if he really makes 80% in the long run. Based on the simulation, our estimate of this probability is 3/400 = 0.0075. The observed statistic, $\hat{p} = 0.64$, is so unlikely if the actual parameter value is $p = 0.80$ that it gives convincing evidence that the player's claim is not true.

Be sure that you understand why this evidence is convincing. There are two possible explanations of the fact that our virtual player made only $\hat{p} = 32/50 = 0.64$ of his free throws:

1. The null hypothesis is correct ($p = 0.8$), and just by chance, a very unlikely outcome occurred.
2. The alternative hypothesis is correct—the population proportion is less than 0.8, so the sample result is not an unlikely outcome.

Explanation 1 might be correct—the result of our random sample of 50 shots *could* be due to chance alone. But the probability that such a result would occur by chance is so small (less than 1 in a 100) that we are quite confident that Explanation 2 is right.

Statistical tests use an elaborate vocabulary, but the basic idea is: *an outcome that would rarely happen if the null hypothesis were true is good evidence that the null hypothesis is not true.*

Interpreting *P*-Values

The idea of stating a null hypothesis that we want to find evidence *against* seems odd at first. It may help to think of a criminal trial. The defendant is "innocent until proven guilty." That is, the null hypothesis is innocence and the prosecution must try to provide convincing evidence against this hypothesis and in favor of the alternative hypothesis: guilt. That's exactly how statistical tests work, although in statistics we deal with evidence provided by data and use a probability to say how strong the evidence is.

The null hypothesis H_0 states the claim that we are seeking evidence against. The probability that measures the strength of the evidence against H_0 and in favor of H_a is called a **P-value**.

DEFINITION: *P*-value

The probability, computed assuming H_0 is true, that the statistic (such as $\hat{p}$ or $\bar{x}$) would take a value as extreme as or more extreme than the one actually observed, in the direction specified by H_a, is called the ***P*-value** of the test.

Small *P*-values are evidence against H_0 because they say that the observed result is unlikely to occur when H_0 is true. Large *P*-values fail to give convincing evidence against H_0 and in favor of H_a because they say that the observed result is likely to occur by chance alone when H_0 is true.

We'll show you how to calculate *P*-values later. For now, let's focus on interpreting them.

EXAMPLE I'm a Great Free-Throw Shooter!

Interpreting a P-value

The *P*-value is the probability of getting a sample result at least as extreme as the one we did if H_0 were true. Because the alternative hypothesis is $H_a: p < 0.80$, the sample results that count as "at least as extreme" are those with $\hat{p} \leq 0.64$. In other words, the *P*-value is the conditional probability $P(\hat{p} \leq 0.64 \mid p = 0.80)$. Earlier, we used a simulation to estimate this probability as $3/400 = 0.0075$. So *if H_0 is true and the player makes 80% of his free throws in the long run, there's about a 0.0075 probability that the player would make 32 or fewer of 50 shots by chance alone.* The small probability gives strong evidence against H_0 and in favor of the alternative $H_a: p < 0.80$ because it would be so unlikely for this result to occur just by chance if H_0 were true.

The alternative hypothesis sets the direction that counts as evidence against H_0. In the previous example, only values of $\hat{p}$ that are much less than 0.80 count as evidence against the null hypothesis because the alternative is one-sided on the low side. If the alternative is two-sided, both directions count.

EXAMPLE Healthy Bones

Interpreting a P-value

Calcium is a vital nutrient for healthy bones and teeth. The National Institutes of Health (NIH) recommends a calcium intake of 1300 mg per day for teenagers. The NIH is concerned that teenagers aren't getting enough calcium. Is this true?

Researchers want to perform a test of

$$H_0: \mu = 1300$$
$$H_a: \mu < 1300$$

where μ is the true mean daily calcium intake in the population of teenagers. They ask a random sample of 20 teens to record their food and drink consumption for 1 day. The researchers then compute the calcium intake for each student. Data analysis reveals that $\bar{x} = 1198$ mg and $s_x = 411$ mg. After checking that conditions were met, researchers performed a significance test and obtained a P-value of 0.1404.

PROBLEM:

(a) Explain what it would mean for the null hypothesis to be true in this setting.

(b) Interpret the P-value in context.

SOLUTION:

(a) In this setting, $H_0: \mu = 1300$ says that the mean daily calcium intake in the population of teenagers is 1300 mg. If H_0 is true, then teenagers are getting enough calcium, on average.

(b) Assuming that the mean daily calcium intake in the teen population is 1300 mg, there is a 0.1404 probability of getting a sample mean of 1198 mg or less just by chance in a random sample of 20 teens.

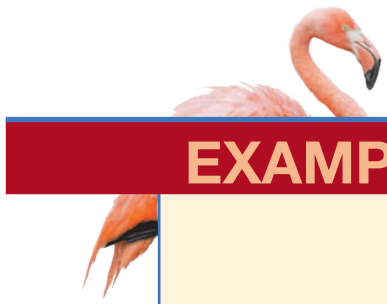
For Practice *Try Exercise* 11

Statistical Significance

The final step in performing a significance test is to draw a conclusion about the competing claims you were testing. We will make one of two decisions based on the strength of the evidence against the null hypothesis (and in favor of the alternative hypothesis)—**reject H_0** or **fail to reject H_0**. If our sample result is too unlikely to have happened by chance assuming H_0 is true, then we'll reject H_0 and say that there is convincing evidence for H_a. Otherwise, we will fail to reject H_0 and say that there is *not* convincing evidence for H_a.

This wording may seem unusual at first, but it's consistent with what happens in a criminal trial. Once the jury has weighed the evidence against the null

hypothesis of innocence, they return one of two verdicts: "guilty" (reject H_0) or "not guilty" (fail to reject H_0). A not-guilty verdict doesn't guarantee that the defendant is innocent, just that there's not convincing evidence of guilt. Likewise, a fail-to-reject H_0 decision in a significance test doesn't mean that H_0 is true. For that reason, *you should never "accept H_0" or use language implying that you believe H_0 is true.*



EXAMPLE Free Throws and Healthy Bones

Drawing conclusions

In the free-throw shooter example, because the estimated *P*-value of 0.0075 is so small, there is strong evidence against the null hypothesis $H_0: p = 0.80$. For that reason, we would reject H_0 in favor of the alternative $H_a: p < 0.80$. We have convincing evidence that the virtual player makes fewer than 80% of his free throws.

For the teen calcium study, however, the large *P*-value of 0.1404 gives weak evidence against $H_0: \mu = 1300$ and in favor of $H_a: \mu < 1300$. We therefore fail to reject H_0. Researchers do not have convincing evidence that teens are getting less than 1300 mg of calcium per day, on average.

In a nutshell, our conclusion in a significance test comes down to

P-value small → reject H_0 → convincing evidence for H_a (in context)

P-value large → fail to reject H_0 → not convincing evidence for H_a (in context)

There is no rule for how small a *P*-value is required to reject H_0—it's a matter of judgment and depends on the specific circumstances. But we can compare the *P*-value with a fixed value that we regard as decisive, called the **significance level**. We write it as α, the Greek letter alpha.

If we choose $\alpha = 0.05$, we are requiring that the data give evidence against H_0 so strong that it would happen less than 5% of the time just by chance when H_0 is true. If we choose $\alpha = 0.01$, we are insisting on stronger evidence against the null hypothesis, a result that would occur less often than 1 in every 100 times by chance alone when H_0 is true.

In Chapter 4, we said that an observed result is "statistically significant" if it would rarely occur by chance alone. When our *P*-value is less than the chosen α in a significance test, we say that the result is **statistically significant at level α**.

DEFINITION: Statistically significant at level α

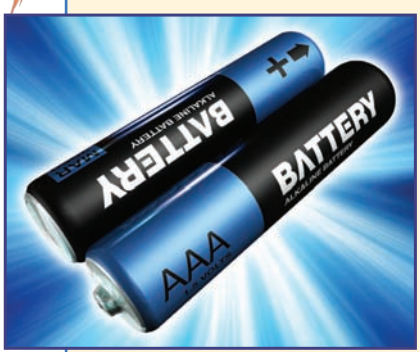
If the *P*-value is smaller than alpha, we say that the results of a study are **statistically significant at level α**. In that case, we reject the null hypothesis H_0 and conclude that there is convincing evidence in favor of the alternative hypothesis H_a.

"Significant" in the statistical sense does not necessarily mean "important." It means simply "not likely to happen just by chance." The significance level α makes "not likely" more exact.

Significance at level 0.01 is often expressed by the statement "The results were significant ($P < 0.01$)." Here, P stands for the P-value. The actual P-value is more informative than a statement of significance because it allows us to assess significance at any level we choose. For example, a result with $P = 0.03$ is significant at the $\alpha = 0.05$ level but is not significant at the $\alpha = 0.01$ level. When we use a fixed significance level to draw a conclusion in a statistical test,

P-value $< \alpha \rightarrow$ reject $H_0 \rightarrow$ convincing evidence for H_a (in context)

P-value $\geq \alpha \rightarrow$ fail to reject $H_0 \rightarrow$ not convincing evidence for H_a (in context)



AP® EXAM TIP The conclusion to a significance test should always include three components: (1) an explicit comparison of the P-value to a stated significance level, (2) a decision about the null hypothesis: reject or fail to reject H_0, and (3) a statement in the context of the problem about whether or not there is convincing evidence for H_a.

EXAMPLE Better Batteries

Statistical significance

A company has developed a new deluxe AAA battery that is supposed to last longer than its regular AAA battery.[2] However, these new batteries are more expensive to produce, so the company would like to be convinced that they really do last longer. Based on years of experience, the company knows that its regular AAA batteries last for 30 hours of continuous use, on average. The company selects an SRS of 15 new batteries and uses them continuously until they are completely drained. The sample mean lifetime is $\bar{x} = 33.9$ hours. A significance test is performed using the hypotheses

$$H_0: \mu = 30 \text{ hours}$$
$$H_a: \mu > 30 \text{ hours}$$

where μ is the true mean lifetime of the new deluxe AAA batteries. The resulting P-value is 0.0729.

PROBLEM: *What conclusion would you make for each of the following significance levels? Justify your answer.*

(a) $\alpha = 0.10$ (b) $\alpha = 0.05$

SOLUTION:

(a) *Because the P-value, 0.0729, is less than $\alpha = 0.10$, we reject H_0. We have convincing evidence that the company's deluxe AAA batteries last longer than 30 hours, on average.*

(b) *Because the P-value, 0.0729, is greater than $\alpha = 0.05$, we fail to reject H_0. We do not have convincing evidence that the company's deluxe AAA batteries last longer than 30 hours, on average.*

For Practice *Try Exercise* 13

In practice, the most commonly used significance level is $\alpha = 0.05$. This is mainly due to Sir Ronald A. Fisher, a famous statistician who worked on agricultural experiments in England during the early twentieth century. Fisher was the first to suggest deliberately using random assignment in an experiment. In a paper published in 1926, Fisher wrote that it is convenient to draw the line at

about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials."[3]

Sometimes it may be preferable to choose $\alpha = 0.01$ or $\alpha = 0.10$, for reasons we will discuss shortly. *Warning:* if you are going to draw a conclusion based on a significance level α, then α should be stated *before* the data are produced. Otherwise, a deceptive user of statistics might set an α level *after* the data have been analyzed in an attempt to manipulate the conclusion. This is just as inappropriate as choosing an alternative hypothesis to be one-sided in a particular direction *after* looking at the data.

THINK ABOUT IT

How do you choose a significance level? The purpose of a significance test is to give a clear statement of the strength of evidence provided by the data against the null hypothesis and in favor of the alternative hypothesis. The *P*-value does this. But how small a *P*-value is convincing evidence against the null hypothesis? This depends mainly on two circumstances:

- *How plausible is* H_0? If H_0 represents an assumption that the people you must convince have believed for years, strong evidence (a very small *P*-value) will be needed to persuade them.
- *What are the consequences of rejecting* H_0? If rejecting H_0 in favor of H_a means making an expensive change of some kind, you need strong evidence that the change will be beneficial.

These criteria are a bit subjective. Different people will insist on different levels of significance. Giving the *P*-value allows each of us to decide individually if the evidence is sufficiently strong.

Users of statistics have often emphasized standard significance levels such as 10%, 5%, and 1%. The 5% level, $\alpha = 0.05$, is very common. For example, courts have tended to accept 5% as a standard in discrimination cases.[4]

Beginning users of statistical tests generally find it easier to compare a *P*-value to a significance level than to interpret the *P*-value correctly in context. For that reason, we will include stating a significance level as a required part of every significance test. We'll also ask you to explain what a *P*-value means in a variety of settings.

Type I and Type II Errors

When we draw a conclusion from a significance test, we hope our conclusion will be correct. But sometimes it will be wrong. There are two types of mistakes we can make. We can reject the null hypothesis when it's actually true, known as a **Type I error,** or we can fail to reject H_0 when the alternative hypothesis is true, which is a **Type II error**.

DEFINITION: Type I error and Type II error

If we reject H_0 when H_0 is true, we have committed a **Type I error**.

If we fail to reject H_0 when H_a is true, we have committed a **Type II error**.

		Truth about the population: H_0 true	Truth about the population: H_a true
Conclusion based on sample	Reject H_0	Type I error	Correct conclusion
	Fail to reject H_0	Correct conclusion	Type II error

FIGURE 9.2 The two types of errors in significance tests.

The possibilities are summarized in Figure 9.2. If H_0 is true, our conclusion is correct if we fail to reject H_0, but it is a Type I error if we reject H_0. If H_a is true, our conclusion is correct if we reject H_0, but is a Type II error if we fail to reject H_0. Only one error is possible at a time.

It is important to be able to describe Type I and Type II errors in the context of a problem. Considering the consequences of each of these types of error is also important, as the following example shows.

EXAMPLE Perfect Potatoes

Type I and Type II errors

A potato chip producer and its main supplier agree that each shipment of potatoes must meet certain quality standards. If the producer determines that more than 8% of the potatoes in the shipment have "blemishes," the truck will be sent away to get another load of potatoes from the supplier. Otherwise, the entire truckload will be used to make potato chips. To make the decision, a supervisor will inspect a random sample of potatoes from the shipment. The producer will then perform a significance test using the hypotheses

$$H_0: p = 0.08$$
$$H_a: p > 0.08$$

where p is the actual proportion of potatoes with blemishes in a given truckload.

PROBLEM: Describe a Type I and a Type II error in this setting, and explain the consequences of each.

SOLUTION: A Type I error occurs if we reject H_0 when H_0 is true. That would happen if the producer finds convincing evidence that the proportion of potatoes with blemishes is greater than 0.08 when the actual proportion is 0.08 (or less). Consequence: The potato-chip producer sends the truckload of acceptable potatoes away, which may result in lost revenue for the supplier. Furthermore, the producer will have to wait for another shipment of potatoes before producing the next batch of potato chips.

A Type II error occurs if we fail to reject H_0 when H_a is true. That would happen if the producer does not find convincing evidence that more than 8% of the potatoes in the shipment have blemishes when that is actually the case. Consequence: The producer uses the truckload of potatoes to make potato chips. More chips will be made with blemished potatoes, which may upset customers.

Here's a helpful reminder to keep the two types of errors straight. "Fail to" goes with Type II.

For Practice *Try Exercise* 21

Which is more serious: a Type I error or a Type II error? That depends on the situation. For the potato-chip producer, a Type II error could result in upset customers, leading to decreased sales. A Type I error, turning away a shipment even though 8% or less of the potatoes have blemishes, may not have much impact if additional shipments of potatoes can be obtained fairly easily. However, the supplier won't be too happy with a Type I error.

CHECK YOUR UNDERSTANDING

Refer to the "Better Batteries" example on page 546.

1. Describe a Type I error in this setting.
2. Describe a Type II error in this setting.
3. Which type of error is more serious in this case? Justify your answer.

Error Probabilities We can assess the performance of a significance test by looking at the probabilities of the two types of error. That's because statistical inference is based on asking, "What would happen if I repeated the data-production process many times?" We cannot (without inspecting the whole truckload) guarantee that good shipments of potatoes will never be sent away and bad shipments will never be used to make chips. But we can think about our chances of making each of these mistakes.

EXAMPLE Perfect Potatoes

Type I error probability

For the truckload of potatoes in the previous example, we were testing

$$H_0: p = 0.08$$
$$H_a: p > 0.08$$

where p is the proportion of all potatoes with blemishes in the shipment. Suppose that the potato-chip producer decides to carry out this test based on a random sample of 500 potatoes using a 5% significance level ($\alpha = 0.05$).

A Type I error is to reject H_0 when H_0 is actually true. If our sample results in a value of $\hat{p}$ that is much larger than 0.08, we will reject H_0. How large would $\hat{p}$ need to be? The 5% significance level tells us to count results that could happen less than 5% of the time by chance if H_0 is true as evidence that H_0 is false.

Assuming $H_0: p = 0.08$ is true, the sampling distribution of $\hat{p}$ will have

Shape: Approximately Normal because $np = 500(0.08) = 40$ and $n(1 - p) = 500(0.92) = 460$ are both at least 10.

Center: $\mu_{\hat{p}} = p = 0.08$

Spread: $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}} = \sqrt{\dfrac{0.08(0.92)}{500}} = 0.01213$, assuming that there are at least $10(500) = 5000$ potatoes in the shipment.

Figure 9.3 on the next page shows the Normal curve that approximates this sampling distribution.

The shaded area in the right tail of Figure 9.3 is 5%. We used the calculator command `invNorm(area: .95, μ: .08, σ: .01213)` to get the boundary value $\hat{p} = 0.10$. Values of $\hat{p}$ to the right of the green line at $\hat{p} = 0.10$ will cause us to reject H_0 even though H_0 is true. This will happen in 5% of all possible samples. That is, the probability of making a Type I error is 0.05.

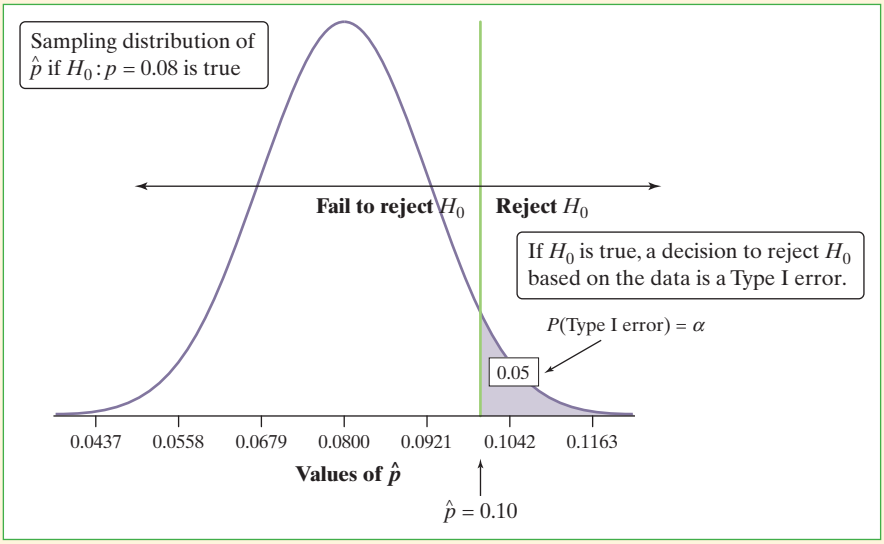


FIGURE 9.3 The probability of a Type I error (shaded area) is the probability of rejecting $H_0: p = 0.08$ when H_0 is actually true.

The probability of a Type I error is the probability of rejecting H_0 when it is true. As the previous example showed, this is exactly the significance level of the test.

SIGNIFICANCE AND TYPE I ERROR

The significance level α of any fixed-level test is the probability of a Type I error. That is, α is the probability that the test will reject the null hypothesis H_0 when H_0 is actually true. Consider the consequences of a Type I error before choosing a significance level.

What about Type II errors? We'll discuss them at the end of Section 9.2, after you have learned how to carry out a significance test.

Section 9.1 Summary

- **A significance test** assesses the evidence provided by data against a **null hypothesis** H_0 and in favor of an **alternative hypothesis** H_a.
- The hypotheses are usually stated in terms of population parameters. Often, H_0 is a statement of no change or no difference. The alternative hypothesis states what we hope or suspect is true.
- A **one-sided alternative** H_a says that a parameter differs from the null hypothesis value in a specific direction. A **two-sided alternative** H_a says that a parameter differs from the null value in either direction.

- The reasoning of a significance test is as follows. Suppose that the null hypothesis is true. If we repeated our data production many times, would we often get data as inconsistent with H_0, in the direction specified by H_a, as the data we actually have? If the data are unlikely when H_0 is true, they provide evidence against H_0 and in favor of H_a.
- The ***P*-value** of a test is the probability, computed supposing H_0 to be true, that the statistic will take a value at least as extreme as the observed result in the direction specified by H_a.
- Small *P*-values indicate strong evidence against H_0. To calculate a *P*-value, we must know the sampling distribution of the test statistic when H_0 is true.
- If the *P*-value is smaller than a specified value α (called the **significance level**), the data are **statistically significant at level α**. In that case, we can **reject** H_0 and say that we have convincing evidence for H_a. If the *P*-value is greater than or equal to α, we **fail to reject** H_0 and say that we do *not* have convincing evidence for H_a.
- A **Type I error** occurs if we reject H_0 when it is in fact true. In other words, the data give convincing evidence for H_a when the null hypothesis is correct. A **Type II error** occurs if we fail to reject H_0 when H_a is true. In other words, the data don't give convincing evidence for H_a, even though the alternative hypothesis is correct.
- In a fixed level α significance test, the probability of a Type I error is the significance level α.

Section 9.1 Exercises

In Exercises 1 to 6, each situation calls for a significance test. State the appropriate null hypothesis H_0 and alternative hypothesis H_a in each case. Be sure to define your parameter each time.

1. **Attitudes** The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures students' attitudes toward school and study habits. Scores range from 0 to 200. The mean score for U.S. college students is about 115. A teacher suspects that older students have better attitudes toward school. She gives the SSHA to an SRS of 45 of the over 1000 students at her college who are at least 30 years of age.

pg 540

2. **Anemia** Hemoglobin is a protein in red blood cells that carries oxygen from the lungs to body tissues. People with less than 12 grams of hemoglobin per deciliter of blood (g/dl) are anemic. A public health official in Jordan suspects that Jordanian children are at risk of anemia. He measures a random sample of 50 children.

3. **Lefties** Simon reads a newspaper report claiming that 12% of all adults in the United States are left-handed. He wonders if the proportion of lefties at his large community college is really 12%. Simon chooses an SRS of 100 students and records whether each student is right- or left-handed.

4. **Don't argue!** A Gallup Poll report revealed that 72% of teens said they seldom or never argue with their friends.[5] Yvonne wonders whether this result holds true in her large high school. So she surveys a random sample of 150 students at her school.

5. **Cold cabin?** During the winter months, the temperatures at the Colorado cabin owned by the Starnes family can stay well below freezing (32°F or 0°C) for weeks at a time. To prevent the pipes from freezing, Mrs. Starnes sets the thermostat at 50°F. The manufacturer claims that the thermostat allows variation in home temperature of $\sigma = 3$°F. Mrs. Starnes suspects that the manufacturer is overstating how well the thermostat works.

6. **Ski jump** When ski jumpers take off, the distance they fly varies considerably depending on their speed, skill, and wind conditions. Event organizers

must position the landing area to allow for differences in the distances that the athletes fly. For a particular competition, the organizers estimate that the variation in distance flown by the athletes will be $\sigma = 10$ meters. An experienced jumper thinks that the organizers are underestimating the variation.

In Exercises 7 to 10, explain what's wrong with the stated hypotheses. Then give correct hypotheses.

7. **Better parking** A change is made that should improve student satisfaction with the parking situation at a local high school. Right now, 37% of students approve of the parking that's provided. The null hypothesis $H_0: p > 0.37$ is tested against the alternative $H_a: p = 0.37$.

8. **Better parking** A change is made that should improve student satisfaction with the parking situation at your school. Right now, 37% of students approve of the parking that's provided. The null hypothesis $H_0: \hat{p} = 0.37$ is tested against the alternative $H_a: \hat{p} \neq 0.37$.

9. **Birth weights** In planning a study of the birth weights of babies whose mothers did not see a doctor before delivery, a researcher states the hypotheses as

$$H_0: \bar{x} = 1000 \text{ grams}$$

$$H_a: \bar{x} < 1000 \text{ grams}$$

10. **Birth weights** In planning a study of the birth weights of babies whose mothers did not see a doctor before delivery, a researcher states the hypotheses as

$$H_0: \mu < 1000 \text{ grams}$$

$$H_a: \mu = 900 \text{ grams}$$

11. **Attitudes** In the study of older students' attitudes from Exercise 1, the sample mean SSHA score was 125.7 and the sample standard deviation was 29.8. A significance test yields a *P*-value of 0.0101. pg 544

(a) Explain what it would mean for the null hypothesis to be true in this setting.

(b) Interpret the *P*-value in context.

12. **Anemia** For the study of Jordanian children in Exercise 2, the sample mean hemoglobin level was 11.3 g/dl and the sample standard deviation was 1.6 g/dl. A significance test yields a *P*-value of 0.0016.

(a) Explain what it would mean for the null hypothesis to be true in this setting.

(b) Interpret the *P*-value in context.

13. **Lefties** Refer to Exercise 3. In Simon's SRS, 16 of the students were left-handed. A significance test yields a *P*-value of 0.2184. What conclusion would you make if $\alpha = 0.10$? If $\alpha = 0.05$? Justify your answers. pg 546

14. **Don't argue!** Refer to Exercise 4. For Yvonne's survey, 96 students in the sample said they rarely or never argue with friends. A significance test yields a *P*-value of 0.0291. What conclusion would you make if $\alpha = 0.05$? If $\alpha = 0.01$? Justify your answers.

15. **Attitudes** Refer to Exercise 11. What conclusion would you make if $\alpha = 0.05$? If $\alpha = 0.01$? Justify your answers.

16. **Anemia** Refer to Exercise 12. What conclusion would you make if $\alpha = 0.05$? If $\alpha = 0.01$? Justify your answers.

17. **Interpreting a *P*-value** When asked to explain the meaning of the *P*-value in Exercise 13, a student says, "This means there is about a 22% chance that the null hypothesis is true." Explain why the student's explanation is wrong.

18. **Interpreting a *P*-value** When asked to explain the meaning of the *P*-value in Exercise 14, a student says, "There is a 0.0291 probability of getting a sample proportion of $\hat{p} = 96/150 = 0.64$ by chance alone." Explain why the student's explanation is wrong.

19. **Drawing conclusions** A student performs a test of $H_0: p = 0.75$ versus $H_a: p > 0.75$ and gets a *P*-value of 0.99. The student writes: "Because the *P*-value is greater than 0.75, we reject H_0. The data prove that H_a is true." Explain what is wrong with this conclusion.

20. **Drawing conclusions** A student performs a test of $H_0: p = 0.5$ versus $H_a: p \neq 0.5$ and gets a *P*-value of 0.63. The student writes: "Because the *P*-value is greater than $\alpha = 0.05$, we accept H_0. The data provide convincing evidence that the null hypothesis is true." Explain what is wrong with this conclusion.

Exercises 21 and 22 refer to the following setting. Slow response times by paramedics, firefighters, and policemen can have serious consequences for accident victims. In the case of life-threatening injuries, victims generally need medical attention within 8 minutes of the accident. Several cities have begun to monitor emergency response times. In one such city, the mean response time to all accidents involving life-threatening injuries last year was $\mu = 6.7$ minutes. Emergency personnel arrived within 8 minutes on 78% of all calls involving life-threatening injuries last year. The city manager shares this information and encourages these first responders to "do better." At the end of the year, the city manager selects an SRS of 400 calls involving life-threatening injuries and examines the response times.

21. **Awful accidents** pg 548

(a) State hypotheses for a significance test to determine whether the average response time has decreased. Be sure to define the parameter of interest.

(b) Describe a Type I error and a Type II error in this setting, and explain the consequences of each.

(c) Which is more serious in this setting: a Type I error or a Type II error? Justify your answer.

22. Awful accidents

(a) State hypotheses for a significance test to determine whether first responders are arriving within 8 minutes of the call more often. Be sure to define the parameter of interest.

(b) Describe a Type I error and a Type II error in this setting and explain the consequences of each.

(c) Which is more serious in this setting: a Type I error or a Type II error? Justify your answer.

23. Opening a restaurant You are thinking about opening a restaurant and are searching for a good location. From research you have done, you know that the mean income of those living near the restaurant must be over $85,000 to support the type of upscale restaurant you wish to open. You decide to take a simple random sample of 50 people living near one potential location. Based on the mean income of this sample, you will decide whether to open a restaurant there.[6]

(a) State appropriate null and alternative hypotheses. Be sure to define your parameter.

(b) Describe a Type I and a Type II error, and explain the consequences of each.

24. Blood pressure screening Your company markets a computerized device for detecting high blood pressure. The device measures an individual's blood pressure once per hour at a randomly selected time throughout a 12-hour period. Then it calculates the mean systolic (top number) pressure for the sample of measurements. Based on the sample results, the device determines whether there is convincing evidence that the individual's actual mean systolic pressure is greater than 130. If so, it recommends that the person seek medical attention.

(a) State appropriate null and alternative hypotheses in this setting. Be sure to define your parameter.

(b) Describe a Type I and a Type II error, and explain the consequences of each.

Multiple choice: Select the best answer for Exercises 25 to 28.

25. Experiments on learning in animals sometimes measure how long it takes mice to find their way through a maze. The mean time is 18 seconds for one particular maze. A researcher thinks that a loud noise will cause the mice to complete the maze faster. She measures how long each of 10 mice takes with a noise as stimulus. The appropriate hypotheses for the significance test are

(a) $H_0: \mu = 18$; $H_a: \mu \neq 18$.

(b) $H_0: \mu = 18$; $H_a: \mu > 18$.

(c) $H_0: \mu < 18$; $H_a: \mu = 18$.

(d) $H_0: \mu = 18$; $H_a: \mu < 18$.

(e) $H_0: \bar{x} = 18$; $H_a: \bar{x} < 18$.

Exercises 26–28 refer to the following setting. Members of the city council want to know if a majority of city residents supports a 1% increase in the sales tax to fund road repairs. To investigate, they survey a random sample of 300 city residents and use the results to test the following hypotheses:

$$H_0: p = 0.50$$
$$H_a: p > 0.50$$

where p is the proportion of all city residents who support a 1% increase in the sales tax to fund road repairs.

26. A Type I error in the context of this study occurs if the city council

(a) finds convincing evidence that a majority of residents supports the tax increase, when in reality there isn't convincing evidence that a majority supports the increase.

(b) finds convincing evidence that a majority of residents supports the tax increase, when in reality at most 50% of city residents support the increase.

(c) finds convincing evidence that a majority of residents supports the tax increase, when in reality more than 50% of city residents do support the increase.

(d) does not find convincing evidence that a majority of residents supports the tax increase, when in reality more than 50% of city residents do support the increase.

(e) does not find convincing evidence that a majority of residents supports the tax increase, when in reality at most 50% of city residents do support the increase.

27. In the sample, $\hat{p} = 158/300 = 0.527$. The resulting P-value is 0.18. What is the correct interpretation of this P-value?

(a) Only 18% of the city residents support the tax increase.

(b) There is an 18% chance that the majority of residents supports the tax increase.

(c) Assuming that 50% of residents support the tax increase, there is an 18% probability that the sample proportion would be 0.527 or higher by chance alone.

(d) Assuming that more than 50% of residents support the tax increase, there is an 18% probability that the sample proportion would be 0.527 or higher by chance alone.

(e) Assuming that 50% of residents support the tax increase, there is an 18% chance that the null hypothesis is true by chance alone.

28. Based on the *P*-value in Exercise 27, which of the following would be the most appropriate conclusion?

(a) Because the *P*-value is large, we reject H_0. We have convincing evidence that more than 50% of city residents support the tax increase.

(b) Because the *P*-value is large, we fail to reject H_0. We have convincing evidence that more than 50% of city residents support the tax increase.

(c) Because the *P*-value is large, we reject H_0. We have convincing evidence that at most 50% of city residents support the tax increase.

(d) Because the *P*-value is large, we fail to reject H_0. We have convincing evidence that at most 50% of city residents support the tax increase.

(e) Because the *P*-value is large, we fail to reject H_0. We do not have convincing evidence that more than 50% of city residents support the tax increase.

29. **Women in math** (5.3) Of the 24,611 degrees in mathematics given by U.S. colleges and universities in a recent year, 70% were bachelor's degrees, 24% were master's degrees, and the rest were doctorates. Moreover, women earned 43% of the bachelor's degrees, 41% of the master's degrees, and 29% of the doctorates.[7]

(a) How many of the mathematics degrees given in this year were earned by women? Justify your answer.

(b) Are the events "degree earned by a woman" and "degree was a bachelor's degree" independent? Justify your answer using appropriate probabilities.

(c) If you choose 2 of the 24,611 mathematics degrees at random, what is the probability that at least 1 of the 2 degrees was earned by a woman? Show your work.

30. **Explaining confidence** (8.2) Here is an explanation from a newspaper concerning one of its opinion polls. Explain what is wrong with the following statement.

For a poll of 1,600 adults, the variation due to sampling error is no more than three percentage points either way. The error margin is said to be valid at the 95 percent confidence level. This means that, if the same questions were repeated in 20 polls, the results of at least 19 surveys would be within three percentage points of the results of this survey.

9.2 Tests about a Population Proportion

WHAT YOU WILL LEARN By the end of the section, you should be able to:

- State and check the Random, 10%, and Large Counts conditions for performing a significance test about a population proportion.
- Perform a significance test about a population proportion.
- Interpret the power of a test and describe what factors affect the power of a test.
- Describe the relationship among the probability of a Type I error (significance level), the probability of a Type II error, and the power of a test.

Confidence intervals and significance tests are based on the sampling distributions of statistics. That is, both use probability to say what would happen if we used the inference method many times. Section 9.1 presented the reasoning of significance tests, including the idea of a *P*-value. In this section, we focus on the details of testing a claim about a population proportion.

Carrying Out a Significance Test

In Section 9.1, we met a virtual basketball player who claimed to make 80% of his free throws. We thought that he might be exaggerating. In an SRS of 50 shots, the player made only 32. His sample proportion of made free throws was therefore

$$\hat{p} = \frac{32}{50} = 0.64$$

This result is much lower than what he claimed. Does it provide *convincing* evidence against the player's claim? To find out, we need to perform a significance test of

$$H_0: p = 0.80$$
$$H_a: p < 0.80$$

where p = the actual proportion of free throws that the shooter makes in the long run.

Conditions In Chapter 8, we introduced conditions that should be met before we construct a confidence interval for an unknown population proportion: Random, 10% when sampling without replacement, and Large Counts. These same conditions must be verified before carrying out a significance test.

The Large Counts condition for proportions requires that both np and $n(1 - p)$ be at least 10. Because we assume H_0 is true when performing a significance test, we use the parameter value specified by the null hypothesis (sometimes called p_0) when checking the Large Counts condition. In this case, the Large Counts condition says that the expected count of successes np_0 and failures $n(1 - p_0)$ are both at least 10.

CONDITIONS FOR PERFORMING A SIGNIFICANCE TEST ABOUT A PROPORTION

- **Random:** The data come from a well-designed random sample or randomized experiment.
  - **10%:** When sampling without replacement, check that $n \leq \frac{1}{10}N$.
- **Large Counts:** Both np_0 and $n(1 - p_0)$ are at least 10.

Here's an example that shows how to check the conditions.

EXAMPLE I'm a Great Free-Throw Shooter!

Checking conditions

PROBLEM: Check the conditions for performing a significance test of the virtual basketball player's claim.

SOLUTION: The required conditions are

- *Random:* We can view this set of 50 computer-generated shots as a simple random sample from the population of all possible shots that the virtual shooter takes.
  - 10%: We're not sampling without replacement from a finite population (because the applet can keep on shooting), so we don't need to check the 10% condition. (Note that the outcomes of individual shots are independent because they are determined by the computer's random number generator.)
- *Large Counts:* Assuming H_0 is true, $p = 0.80$. Then $np_0 = (50)(0.80) = 40$ and $n(1 - p_0) = (50)(0.20) = 10$ are both at least 10, so this condition is met.

For Practice *Try Exercise* 31

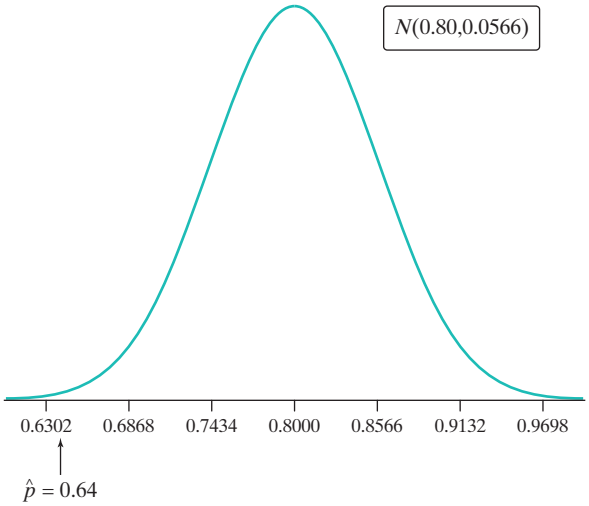


FIGURE 9.4 Normal approximation to the sampling distribution of the proportion $\hat{p}$ of made shots in random samples of 50 free throws by an 80% shooter.

If the null hypothesis $H_0: p = 0.80$ is true, then the player's sample proportion $\hat{p}$ of made free throws in an SRS of 50 shots would vary according to an approximately Normal sampling distribution with mean $\mu_{\hat{p}} = p = 0.80$ and standard deviation

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.80)(0.20)}{50}} = 0.0566.$$

Figure 9.4 displays this distribution. We have added the player's sample result, $\hat{p} = \frac{32}{50} = 0.64$.

Calculations: Test Statistic and *P*-Value A significance test uses sample data to measure the strength of evidence against H_0 and in favor of H_a. Here are some principles that apply to most tests:

- The test compares a statistic calculated from sample data with the value of the parameter stated by the null hypothesis.
- Values of the statistic far from the null parameter value in the direction specified by the alternative hypothesis give strong evidence against H_0.
- To assess *how far* the statistic is from the parameter, standardize the statistic. This standardized value is called the **test statistic**:

On the AP® exam formula sheet, this value is referred to as the "standardized test statistic."

$$\text{test statistic} = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$$

DEFINITION: Test statistic

A **test statistic** measures how far a sample statistic diverges from what we would expect if the null hypothesis H_0 were true, in standardized units. That is,

$$\text{test statistic} = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$$

The test statistic says how far the sample result is from the null parameter value, and in what direction, on a standardized scale. You can use the test statistic to find the *P*-value of the test, as the following example shows.

EXAMPLE I'm a Great Free-Throw Shooter!

Computing the test statistic

PROBLEM: In an SRS of 50 free throws, the virtual player made 32.

(a) Calculate the test statistic.

(b) Find the *P*-value using Table A or technology. Show this result as an area under a standard Normal curve.

SOLUTION:

(a) His sample proportion of made shots is $\hat{p} = 0.64$. Standardizing, we get

$$\text{test statistic} = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$$

$$z = \frac{0.64 - 0.80}{\sqrt{\frac{(0.80)(0.20)}{50}}} = \frac{0.64 - 0.80}{0.0566} = -2.83$$

(b) The shaded area under the curve in Figure 9.5(a) shows the *P*-value. Figure 9.5(b) shows the corresponding area on the standard Normal curve, which displays the distribution of the *z* test statistic. From Table A, we find that the *P*-value is $P(Z \leq -2.83) = 0.0023$.

Using technology: The command `normalcdf (lower: −10000, upper: −2.83, μ : 0, σ : 1)` also gives a *P*-value of 0.0023.

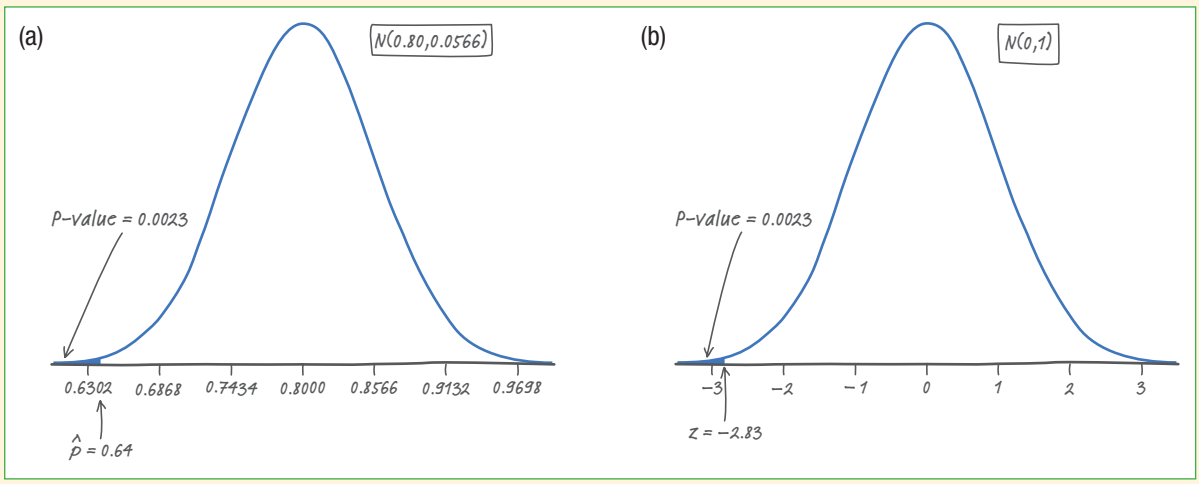


FIGURE 9.5 The shaded area shows the *P*-value for the player's sample proportion of made shots (a) on the Normal approximation to the sampling distribution of $\hat{p}$ from Figure 9.4 and (b) on the standard Normal curve.

If H_0 is true and the player makes 80% of his free throws in the long run, there's only about a 0.0023 probability that he would make 32 or fewer of 50 shots by chance alone.

For Practice *Try Exercise* 35

Earlier, we used simulation to estimate the *P*-value as 0.0075. As the example shows, the *P*-value is even smaller, 0.0023. So if H_0 is true, and the player makes 80% of his free throws in the long run, there's only about a 0.0023 probability that the player would make 32 or fewer of 50 shots by chance alone. This result confirms our earlier decision to reject H_0 and gives convincing evidence that the player is exaggerating.

The One-Sample *z* Test for a Proportion

To perform a significance test, we state hypotheses, check conditions, calculate a test statistic and *P*-value, and draw a conclusion in the context of the problem. The four-step process is ideal for organizing our work.

SIGNIFICANCE TESTS: A FOUR-STEP PROCESS

STEP 4

State: What *hypotheses* do you want to test, and at what *significance level*? Define any *parameters* you use.

Plan: Choose the appropriate inference *method*. Check *conditions*.

Do: If the conditions are met, perform *calculations*.

- Compute the **test statistic**.
- Find the ***P*-value**.

Conclude: Make a *decision* about the hypotheses in the context of the problem.

When the conditions are met—Random, 10%, and Large Counts—the sampling distribution of $\hat{p}$ is approximately Normal with

$$\text{mean } \mu_{\hat{p}} = p \quad \text{and} \quad \text{standard deviation } \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

For confidence intervals, we substitute $\hat{p}$ for p in the standard deviation formula to obtain the standard error. When performing a significance test, however, the null hypothesis specifies a value for p, which we call p_0. We assume that this value is correct when performing our calculations.

If we standardize the statistic $\hat{p}$ by subtracting its mean and dividing by its standard deviation, we get the test statistic:

$$\text{test statistic} = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

This z statistic has approximately the standard Normal distribution when H_0 is true and the conditions are met. *P*-values therefore come from the standard Normal distribution.

Here is a summary of the details for a **one-sample *z* test for a proportion**.

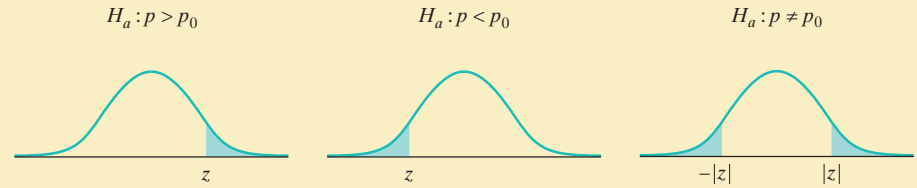
The AP® Statistics course outline calls this test a *large-sample test for a proportion* because it is based on a Normal approximation to the sampling distribution of $\hat{p}$ that becomes more accurate as the sample size increases.

ONE-SAMPLE *z* TEST FOR A PROPORTION

Suppose the conditions are met. To test the hypothesis $H_0: p = p_0$, compute the z statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Find the *P*-value by calculating the probability of getting a z statistic this large or larger in the direction specified by the alternative hypothesis H_a:

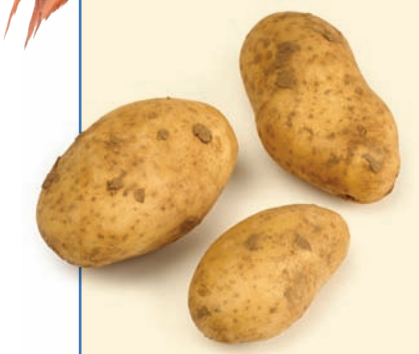


Here is an example of the test in action.

EXAMPLE One Potato, Two Potato

STEP 4

Performing a significance test about p



The potato-chip producer of Section 9.1 has just received a truckload of potatoes from its main supplier. Recall that if the producer finds convincing evidence that more than 8% of the potatoes in the shipment have blemishes, the truck will be sent away to get another load from the supplier. A supervisor selects a random sample of 500 potatoes from the truck. An inspection reveals that 47 of the potatoes have blemishes. Carry out a significance test at the $\alpha = 0.05$ significance level. What should the producer conclude?

STATE: We want to perform a test of

$$H_0: p = 0.08$$
$$H_a: p > 0.08$$

where p is the actual proportion of potatoes in this shipment with blemishes. We'll use an $\alpha = 0.05$ significance level.

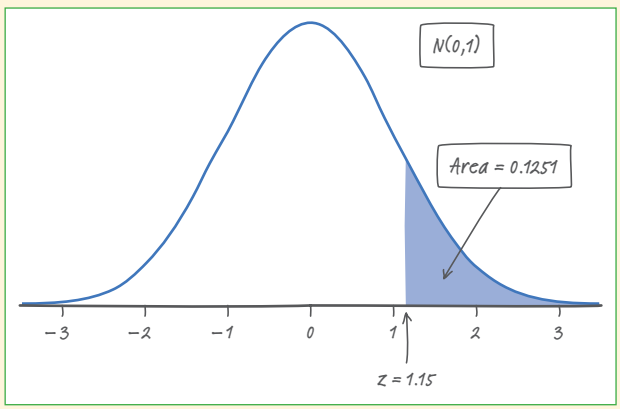


FIGURE 9.6 The *P*-value for the one-sided test.

PLAN: If conditions are met, we should do a one-sample z test for the population proportion p.

- *Random:* The supervisor took a random sample of 500 potatoes from the shipment.
  - *10%:* It seems reasonable to assume that there are at least $10(500) = 5000$ potatoes in the shipment.
- *Large Counts:* Assuming $H_0: p = 0.08$ is true, the expected counts of blemished and unblemished potatoes are $np_0 = 500(0.08) = 40$ and $n(1 - p_0) = 500(0.92) = 460$, respectively. Because both of these values are at least 10, we should be safe doing Normal calculations.

DO: The sample proportion of blemished potatoes is $\hat{p} = 47/500 = 0.094$

- Test statistic $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} = \dfrac{0.094 - 0.08}{\sqrt{\dfrac{0.08(0.92)}{500}}} = 1.15$

- *P*-value Figure 9.6 displays the *P*-value as an area under the standard Normal curve for this one-sided test. Table A gives the *P*-value as $P(Z \geq 1.15) = 1 - 0.8749 = 0.1251$.
- *Using technology:* The command `normalcdf(lower:1.15, upper:10000, μ:0, σ:1)` also gives a *P*-value of 0.1251.

CONCLUDE: Because our *P*-value, 0.1251, is greater than $\alpha = 0.05$, we fail to reject H_0. There is not convincing evidence that the shipment contains more than 8% blemished potatoes. As a result, the producer will use this truckload of potatoes to make potato chips.

AP® EXAM TIP When a significance test leads to a fail to reject H_0 decision, as in this example, be sure to interpret the results as "we don't have convincing evidence to conclude H_a." Saying anything that sounds like you believe H_0 is (or might be) true will lead to a loss of credit. And don't write text-message-type responses, like "FTR the H_0."

For Practice *Try Exercise* 39

The preceding example reminds us why significance tests are important. The sample proportion of blemished potatoes was $\hat{p} = 47/500 = 0.094$. This result gave evidence against H_0 in favor of H_a. To see whether such an outcome is unlikely to occur just by chance when H_0 is true, we had to carry out a significance test. The *P*-value told us that a sample proportion this large or larger would occur in about 12.5% of all random samples of 500 potatoes when H_0 is true. So we can't rule out sampling variability as a plausible explanation for getting a sample proportion of $\hat{p} = 0.094$.

THINK ABOUT IT

What happens when the data don't support H_a? Suppose the supervisor had inspected a random sample of 500 potatoes from the shipment and found 33 with blemishes. This yields a sample proportion of $\hat{p} = 33/500 = 0.066$. Such a sample doesn't give any evidence to support the alternative hypothesis $H_a: p > 0.08$! There's no need to continue with the significance test. The conclusion is clear: we should fail to reject $H_0: p = 0.08$. This truckload of potatoes will be used by the potato-chip producer.

If you weren't paying attention, you might end up carrying out the test. Let's see what would happen. The corresponding test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.066 - 0.08}{\sqrt{\frac{0.08(0.92)}{500}}} = -1.15$$

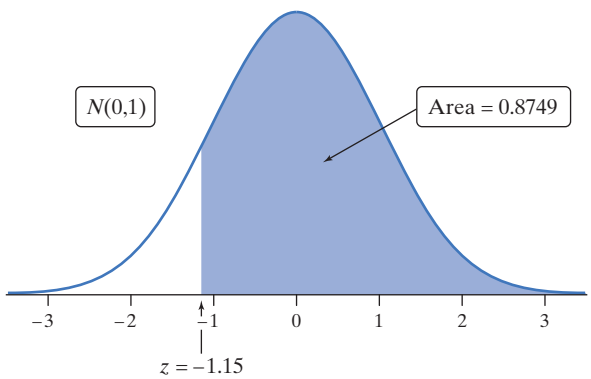


FIGURE 9.7 The *P*-value for the one-sided test.

What's the *P*-value? It's the probability of getting a *z* statistic this large or larger in the direction specified by H_a, $P(Z \geq -1.15)$. Figure 9.7 shows this *P*-value as an area under the standard Normal curve. Using Table A or technology, the *P*-value is $1 - 0.1251 = 0.8749$. There's about an 87.5% chance of getting a sample proportion as large as or larger than $\hat{p} = 0.066$ if $p = 0.08$. As a result, we would fail to reject H_0. Same conclusion, but with lots of unnecessary work!

Always check to see whether the data give evidence against H_0 in the direction specified by H_a before you do calculations.

CHECK YOUR UNDERSTANDING

According to the National Campaign to Prevent Teen and Unplanned Pregnancy, 20% of teens aged 13 to 19 say that they have electronically sent or posted sexually suggestive images of themselves.[8] The counselor at a large high school worries that the actual figure might be higher at her school. To find out, she administers an anonymous survey to a random sample of 250 of the school's 2800 students. All 250 respond, and 63 admit to sending or posting sexual images. Carry out a significance test at the $\alpha = 0.05$ significance level. What conclusion should the counselor draw?

Your calculator will handle the "Do" part of the four-step process, as the following Technology Corner illustrates. However, be sure to read the AP® Exam Tip on the next page.

18. TECHNOLOGY CORNER ONE-PROPORTION *z* TEST ON THE CALCULATOR

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

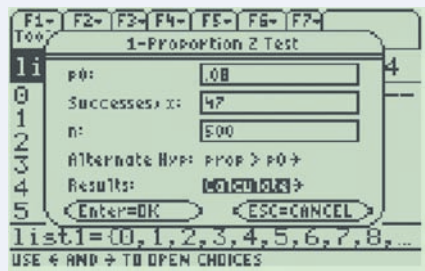
The TI-83/84 and TI-89 can be used to test a claim about a population proportion. We'll demonstrate using the previous example. In a random sample of size $n = 500$, the supervisor found X = 47 potatoes with blemishes. To perform a significance test:

TI-83/84

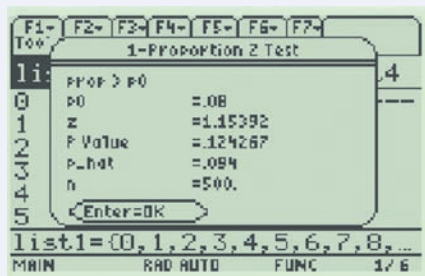
- Press STAT, then choose TESTS and 1-PropZTest.
- On the 1-PropZTest screen, enter the values shown: $p_0 = 0.08$, $x = 47$, and $n = 500$. Specify the alternative hypothesis as "prop > p_0." *Note:* x is the number of successes and n is the number of trials. Both must be whole numbers!

TI-89

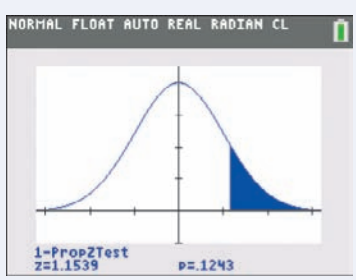
- In the Statistics/List Editor, press 2nd F1 ([F6]) and choose 1-PropZTest.

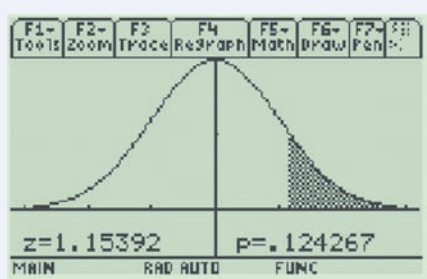


- If you select "Calculate" and press ENTER, you will see that the test statistic is $z = 1.15$ and the P-value is 0.1243.



- If you select the "Draw" option, you will see the screen shown here. Compare these results with those in the example on page 559.





AP® EXAM TIP You can use your calculator to carry out the mechanics of a significance test on the AP® exam. But there's a risk involved. If you just give the calculator answer with no work, and one or more of your values is incorrect, you will probably get no credit for the "Do" step. If you opt for the calculator-only method, be sure to name the procedure (one-proportion z test) and to report the test statistic ($z = 1.15$) and P-value (0.1243).

Two-Sided Tests

Both the free-throw shooter and blemished-potato examples involved one-sided tests. The *P*-value in a one-sided test is the area in one tail of a standard Normal distribution—the tail specified by H_a. In a two-sided test, the alternative hypothesis has the form $H_a: p \neq p_0$. The *P*-value in such a test is the probability of getting a sample proportion as far as or farther from p_0 *in either direction* than the observed value of $\hat{p}$. As a result, you have to find the area in both tails of a standard Normal distribution to get the *P*-value. The following example shows how this process works.

EXAMPLE

Nonsmokers

STEP 4

A two-sided test

According to the Centers for Disease Control and Prevention (CDC) Web site, 50% of high school students have never smoked a cigarette. Taeyeon wonders whether this national result holds true in his large, urban high school. For his AP® Statistics class project, Taeyeon surveys an SRS of 150 students from his school. He gets responses from all 150 students, and 90 say that they have never smoked a cigarette. What should Taeyeon conclude? Give appropriate evidence to support your answer.

STATE: We want to perform a significance test using the hypotheses

$$H_0: p = 0.50$$
$$H_a: p \neq 0.50$$

where p = the proportion of all students in Taeyeon's school who would say they have never smoked cigarettes. Because no significance level was stated, we will use $\alpha = 0.05$.

PLAN: If conditions are met, we'll do a one-sample *z* test for the population proportion *p*.

- *Random:* Taeyeon surveyed an SRS of 150 students from his school.
  - *10%:* It seems reasonable to assume that there are at least $10(150) = 1500$ students in a large high school.
- *Large Counts:* Assuming $H_0: p = 0.50$ is true, the expected counts of smokers and nonsmokers in the sample are $np_0 = 150(0.50) = 75$ and $n(1 - p_0) = 150(0.50) = 75$. Because both of these values are at least 10, we should be safe doing Normal calculations.

DO: The sample proportion is $\hat{p} = 90/150 = 0.60$.

- *Test statistic*

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.60 - 0.50}{\sqrt{\frac{0.50(0.50)}{150}}} = 2.45$$

- *P-value* Figure 9.8 displays the *P*-value as an area under the standard Normal curve for this two-sided test. To compute this *P*-value, we find the area in one tail and double it. Table A gives $P(z \geq 2.45) = 0.0071$ (the right-tail area). So the desired *P*-value is $2(0.0071) = 0.0142$.

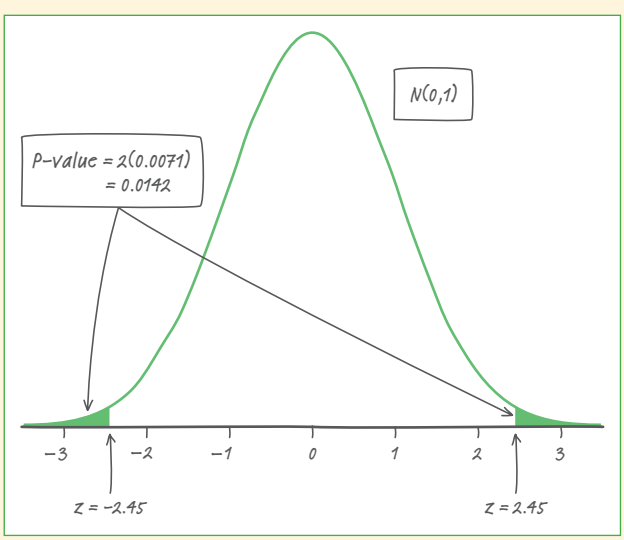


FIGURE 9.8 The *P*-value for the two-sided test.

Using technology: The calculator's 1-PropZTest gives $z = 2.449$ and P-value $= 0.0143$.

CONCLUDE: Because our P-value, 0.0143, is less than $\alpha = 0.05$, we reject H_0. We have convincing evidence that the proportion of all students at Taeyeon's school who would say they have never smoked differs from the national result of 0.50.

For Practice *Try Exercise* 45

CHECK YOUR UNDERSTANDING

According to the National Institute for Occupational Safety and Health, job stress poses a major threat to the health of workers. A news report claims that 75% of restaurant employees feel that work stress has a negative impact on their personal lives.[9] Managers of a large restaurant chain wonder whether this claim is valid for their employees. A random sample of 100 employees finds that 68 answer "Yes" when asked, "Does work stress have a negative impact on your personal life?" Is this good reason to think that the proportion of all employees in this chain who would say "Yes" differs from 0.75? Support your answer with a significance test.

Why Confidence Intervals Give More Information

The result of a significance test begins with a decision to reject H_0 or fail to reject H_0. In Taeyeon's smoking study, for instance, the data led us to reject $H_0: p = 0.50$ because we found convincing evidence that the proportion of students at his school who would say they have never smoked cigarettes differs from the national value. We're left wondering what the actual proportion p might be. A confidence interval might shed some light on this issue.

EXAMPLE

Nonsmokers

A confidence interval gives more info

Taeyeon found that 90 of an SRS of 150 students said that they had never smoked a cigarette. We checked the conditions for performing the significance test earlier. Before we construct a confidence interval for the population proportion p, we should check that both $n\hat{p}$ and $n(1 - \hat{p})$ are at least 10. Because the number of successes and the number of failures in the sample are 90 and 60, respectively, we can proceed with calculations.

Our 95% confidence interval is

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.60 \pm 1.96\sqrt{\frac{0.60(0.40)}{150}} = 0.60 \pm 0.078 = (0.522, 0.678)$$

We are 95% confident that the interval from 0.522 to 0.678 captures the true proportion of students at Taeyeon's high school who would say that they have never smoked a cigarette.

The confidence interval in this example is much more informative than the significance test we performed earlier. The interval gives the values of p that are plausible based on the sample data. We would not be surprised if the true proportion of students at Taeyeon's school who would say they have never smoked cigarettes was as low as 0.522 or as high as 0.678. However, we would be surprised if the true proportion was 0.50 because this value is not contained in the confidence interval. Figure 9.9 gives computer output from Minitab software that includes both the results of the significance test and the confidence interval.

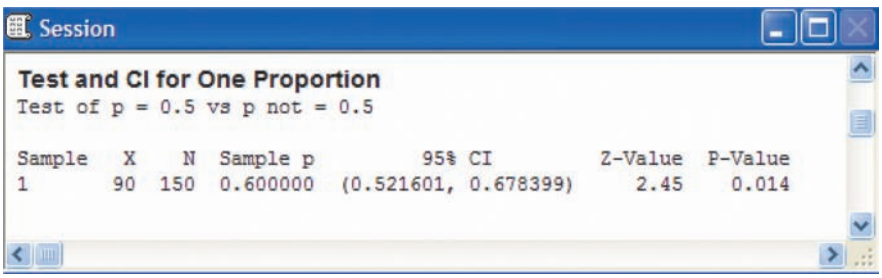


FIGURE 9.9 Minitab output for the two-sided significance test at $\alpha = 0.05$ and a 95% confidence interval for Taeyeon's smoking study.

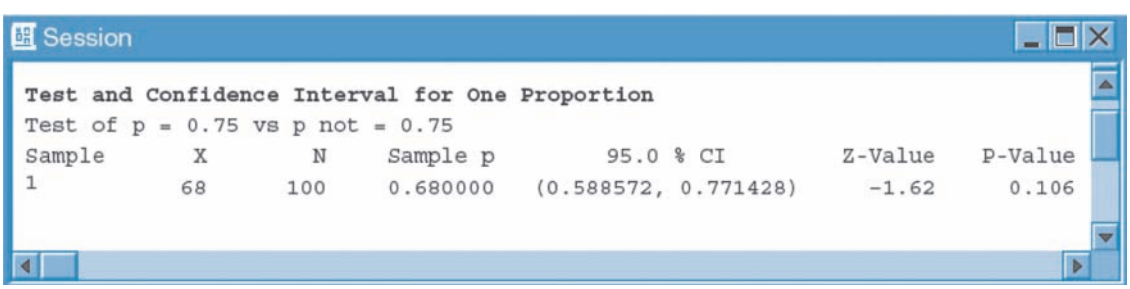
There is a link between confidence intervals and two-sided tests. The 95% confidence interval (0.522, 0.678) gives an approximate set of p_0's that would not be rejected by a two-sided test at the $\alpha = 0.05$ significance level. With proportions, the link isn't perfect because the standard error used for the confidence interval is based on the sample proportion $\hat{p}$, while the denominator of the test statistic is based on the value p_0 from the null hypothesis.

$$\text{Test statistic: } z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \qquad \text{Confidence interval: } \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The big idea is still worth considering: a two-sided test at significance level α and a $100(1-\alpha)\%$ confidence interval (a 95% confidence interval if $\alpha = 0.05$) give similar information about the population parameter.

CHECK YOUR UNDERSTANDING

The figure below shows Minitab output from a significance test and confidence interval for the restaurant worker data in the previous Check Your Understanding (page 563). Explain how the confidence interval is consistent with, but gives more information than, the test.



Type II Error and the Power of a Test

A significance test makes a Type II error when it fails to reject a null hypothesis H_0 that really is false. There are many values of the parameter that make the alternative hypothesis H_a true, so we concentrate on one value. Consider the potato-chip example on page 559 that involved a test of $H_0: p = 0.08$ versus $H_a: p > 0.08$. If the true proportion of blemished potatoes in the shipment was $p = 0.09$, we made a Type II error by failing to reject H_0 based on the sample data. Of course, we also made a Type II error if $p = 0.11$ or $p = 0.15$.

The probability of making a Type II error depends on several factors, including the actual value of the parameter. In the potato-chip example, our test will be more likely to reject $H_0: p = 0.08$ in favor of $H_a: p > 0.08$ if the true proportion of blemished potatoes in the shipment is $p = 0.11$ than if it is $p = 0.09$. Why? because $p = 0.11$ is farther away from the null value than is $p = 0.09$. So we will be less likely to make a Type II error if 11% of potatoes in the shipment are blemished than if only 9% are blemished. A high probability of Type II error for a specific alternative parameter value means that the test is not sensitive enough to usually detect that alternative.

It is more common to report the probability that a significance test *does* reject H_0 when an alternative parameter value is true. This probability is called the **power** of the test against that specific alternative.

DEFINITION: Power

The **power** of a test against a specific alternative is the probability that the test will reject H_0 at a chosen significance level α when the specified alternative value of the parameter is true.

As the following example illustrates, Type II error and power are closely linked.

EXAMPLE

Perfect Potatoes

Type II error and power

The potato-chip producer wonders whether the significance test of $H_0: p = 0.08$ versus $H_a: p > 0.08$ based on a random sample of 500 potatoes has enough power to detect a shipment with, say, 11% blemished potatoes. In this case, a particular Type II error is to fail to reject $H_0: p = 0.08$ when $p = 0.11$. Figure 9.10 on the next page shows two sampling distributions of $\hat{p}$, one when $p = 0.08$ and the other when $p = 0.11$.

Earlier, we decided to reject H_0 if our sample yielded a value of $\hat{p}$ to the right of the green line at $\hat{p} = 0.10$. That decision was based on using a significance level (Type I error probability) of $\alpha = 0.05$.

Now look at the sampling distribution for $p = 0.11$. The shaded area to the right of the green line represents the probability of correctly rejecting H_0 when $p = 0.11$. That is, the *power* of this test to detect $p = 0.11$ is about 0.76. In other words, the potato-chip producer has roughly a 3-in-4 chance of rejecting a truckload with 11% blemished potatoes based on a random sample of 500 potatoes from the shipment.

We would fail to reject H_0 if the sample proportion $\hat{p}$ falls to the left of the green line. The white area under the bottom Normal distribution shows the probability

of failing to reject H_0 when H_0 is false. This is the probability of a Type II error. The potato-chip producer has about a 1-in-4 chance of failing to send away a shipment with 11% blemished potatoes.

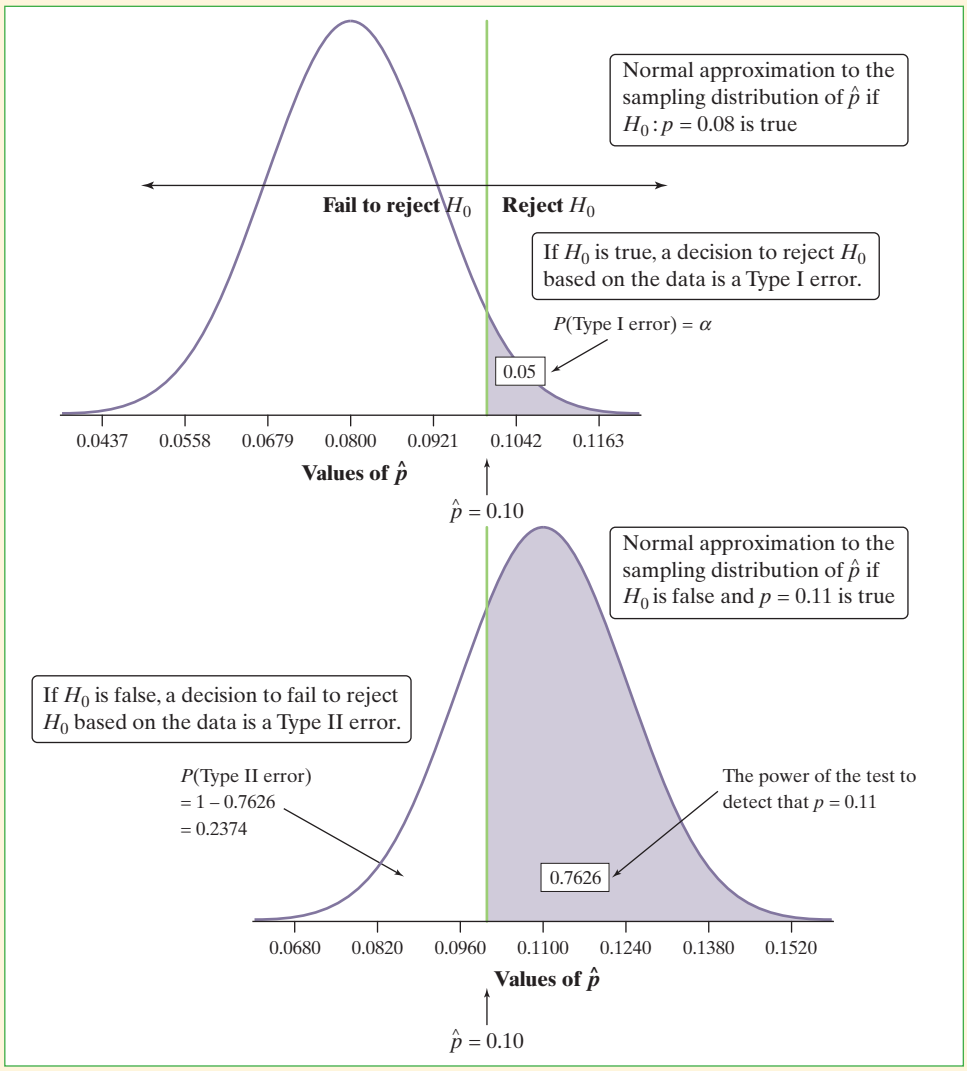


FIGURE 9.10 In the bottom graph, the power of the test (shaded area) is the probability that it correctly rejects $H_0: p = 0.08$ when the truth is $p = 0.11$. In this case, power = 0.7626. The probability of making a Type II error (white area) is $1 - 0.7626 = 0.2374$.

After reading the example, you might be wondering whether 0.76 is a high power or a low power. That depends on how certain the potato-chip producer wants to be to detect a shipment with 11% blemished potatoes. The power of a test against a specific alternative value of the parameter (like $p = 0.11$) is a number between 0 and 1. A power close to 0 means the test has almost no chance of correctly detecting that H_0 is false. A power near 1 means the test is very likely to reject H_0 in favor of H_a when H_0 is false.

The significance level of a test is the probability of reaching the *wrong* conclusion when the null hypothesis is true. The power of a test to detect a specific alternative is the probability of reaching the *right* conclusion when that alternative is true. We can just as easily describe the test by giving the probability of making a Type II error (sometimes called β).

POWER AND TYPE II ERROR

The power of a test against any alternative is 1 minus the probability of a Type II error for that alternative; that is, power $= 1 - \beta$.

Calculating a Type II error probability or power by hand is possible but unpleasant. It's better to let technology do the work for you.

ACTIVITY What Affects the Power of a Test?

MATERIALS:
Computer with Internet access and projection capability

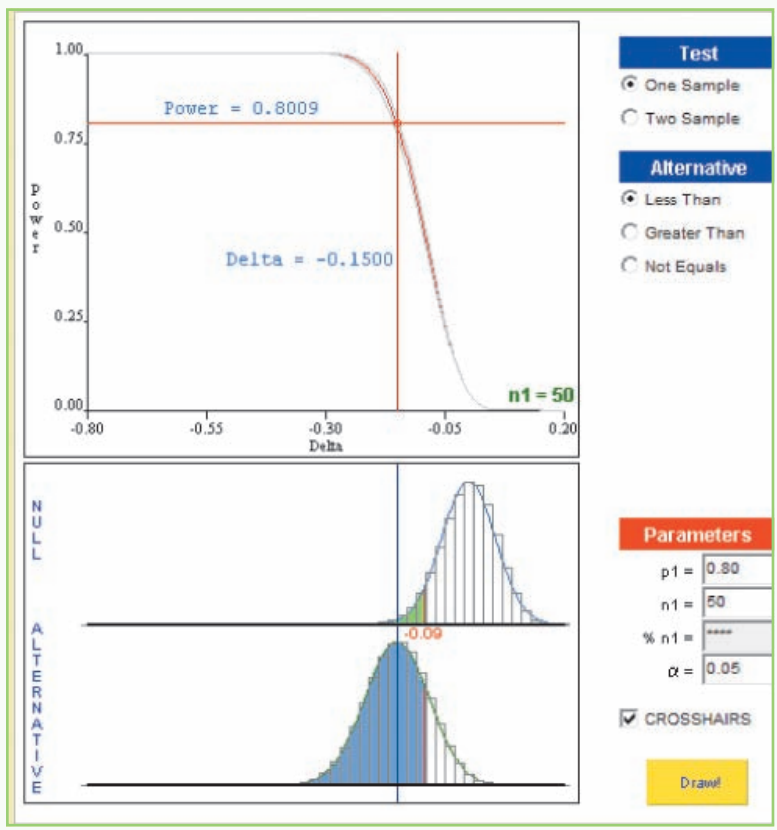
A virtual basketball player claims to make 80% of his free throws. Suppose that the player is exaggerating—he really makes less than 80% in the long run. We have the computer player shoot 50 shots and record the sample proportion $\hat{p}$ of made free throws. We then use the sample result to perform a test of

$$H_0: p = 0.80$$
$$H_a: p < 0.80$$

at the $\alpha = 0.05$ significance level. How can we increase the power of the test to detect that the player is exaggerating? In this Activity, we will use an applet to investigate.

1. Go to www.amstat.org/publications/jse/v11n3/java/Power and select the *Proportions* applet at the bottom of the page.

2. Adjust the applet settings as follows: choose "One Sample" for the Test, "Less Than" for the Alternative, enter 0.8 for p1, 50 for n1, and 0.05 for α. The Null distribution should appear in the applet window.



3. Let's assume for now that the virtual player really makes 65% of his free throws ($p = 0.65$). Drag your mouse to the left in the applet screen and watch as an Alternative distribution appears. Keep dragging until the Delta value in the top panel shows −0.1500. This sets the alternative parameter value to be 0.15 less than the null parameter value of 0.80. Click your mouse to set the Alternative distribution. The Power of the test to detect $p = 0.65$ is shown in the top panel: 0.8009.

4. *Sample size* Change the sample size from $n = 50$ shots to $n = 100$ shots. What happens in the bottom panel of the applet? Does the power increase or decrease? Explain why this makes sense.

5. *Significance level* Reset the sample size to $n = 50$.

(a) Change the significance level to $\alpha = 0.01$. What happens in the bottom panel of the applet? Does the power increase or decrease?

(b) Make a guess about what will happen if you change the significance level to $\alpha = 0.10$. Use the applet to test your conjecture.

(c) Explain what the results in parts (a) and (b) tell you about the relationship between Type I error probability and Type II error probability.

6. *Difference between null parameter value and alternative parameter value* Reset the sample size to $n = 50$ and the significance level to $\alpha = 0.05$. Will we be more likely to detect that the player is really a 65% shooter or that he is really a 70% shooter? Use your mouse to adjust the location of the Alternative distribution. How does the power change? Explain why this makes sense.

Step 5 of the Activity reveals an important link between Type I and Type II error probabilities. Because $P(\text{Type I error}) = \alpha$, increasing the significance level increases the chance of making a Type I error. As the applet shows, this change also increases the power of the test. Because $P(\text{Type II error}) = 1 - \text{Power}$, higher power means a smaller chance of making a Type II error. So increasing the Type I error probability α decreases the Type II error probability β. By the same logic, decreasing the chance of a Type I error results in a higher chance of a Type II error.

Let's summarize what the Activity reveals about how to increase the power of a significance test to detect when H_0 is false and H_a is true.

- **Increase the sample size.** As Step 4 of the Activity confirms, we get better information about the virtual player's free-throw shooting from a random sample of 100 shots than from a random sample of 50 shots. Increasing the sample size decreases the spread of both the Null and Alternative distributions. This change decreases the amount of overlap between the two distributions, making it easier to detect a difference between the null and alternative parameter values.
- **Increase the significance level α.** Using a larger value of α increases the area of the green and blue "reject H_0" regions in both the Null and Alternative distributions. This change makes it more likely to get a sample proportion that leads us to correctly reject the null hypothesis when the shooter is exaggerating.
- **Increase the difference between the null and alternative parameter values that is important to detect.** Step 6 of the Activity shows that it is easier to detect large differences between the null and alternative parameter values than smaller differences. The size of difference that is important to detect is usually determined by experts in the field, so the statistician usually gets little or no input on this factor.

Many researchers who design statistical studies refer to the difference that's important to detect as the *effect size*.

In addition to these three factors, we can also gain power by making wise choices when collecting data. For example, using blocking in an experiment or stratified random sampling can greatly increase the power of a test in some circumstances. Our best advice for maximizing the power of a test is to choose as high an α level (Type I error probability) as you are willing to risk and as large a sample size as you can afford.

CHECK YOUR UNDERSTANDING

Refer to the Perfect Potatoes example on page 565.

1. Which is more serious for the potato-chip producer in this setting: a Type I error or a Type II error? Based on your answer, would you choose a significance level of $\alpha = 0.01$, 0.05, or 0.10?

2. Tell if each of the following would increase or decrease the power of the test. Justify your answers.

(a) Change the significance level to $\alpha = 0.10$.

(b) Take a random sample of 250 potatoes instead of 500 potatoes.

(c) Insist on being able to detect that $p = 0.10$ instead of $p = 0.11$.

Section 9.2 Summary

- The conditions for performing a significance test of $H_0: p = p_0$ are:
  - **Random:** The data were produced by a well-designed random sample or randomized experiment.
    - **10%:** When sampling without replacement, check that the population is at least 10 times as large as the sample.
  - **Large Counts:** The sample is large enough to satisfy $np_0 \geq 10$ and $n(1 - p_0) \geq 10$ (that is, the expected counts of successes and failures are both at least 10).
- The **one-sample *z* test for a population proportion** is based on the **test statistic**

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$$

with *P*-values calculated from the standard Normal distribution.

- Follow the four-step process when you are asked to carry out a significance test:

STATE: What *hypotheses* do you want to test, and at what *significance level*? Define any *parameters* you use.

PLAN: Choose the appropriate inference *method*. Check *conditions*.

DO: If the conditions are met, perform *calculations*.

- Compute the **test statistic**.
- Find the ***P*-value**.

CONCLUDE: Make a *decision* about the hypotheses in the context of the problem.

- Confidence intervals provide additional information that significance tests do not—namely, a set of plausible values for the true population parameter p. A two-sided test of $H_0: p = p_0$ at significance level α gives roughly the same conclusion as a $100(1 - \alpha)\%$ confidence interval.
- The **power** of a significance test against a specific alternative is the probability that the test will reject H_0 when the alternative is true. Power measures the ability of the test to detect an alternative value of the parameter. For a specific alternative, $P(\text{Type II error}) = 1 - \text{power}$.
- There is an important link between the probabilities of Type I and Type II error in a significance test: as one increases, the other decreases.
- We can increase the power of a significance test by increasing the sample size, increasing the significance level, or increasing the difference that is important to detect between the null and alternative parameter values.

9.2 TECHNOLOGY CORNER

TI-Nspire Instructions in Appendix B; HP Prime instructions on the book's Web site.

18. One-proportion z test on the calculator page 561

Section 9.2 Exercises

In Exercises 31 and 32, check that the conditions for carrying out a one-sample z test for the population proportion p are met.

31. **Home computers** Jason reads a report that says 80% of U.S. high school students have a computer at home. He believes the proportion is smaller than 0.80 at his large rural high school. Jason chooses an SRS of 60 students and records whether they have a computer at home. pg 555

32. **Walking to school** A recent report claimed that 13% of students typically walk to school.[10] DeAnna thinks that the proportion is higher than 0.13 at her large elementary school, so she surveys a random sample of 100 students to find out.

In Exercises 33 and 34, explain why we aren't safe carrying out a one-sample z test for the population proportion p.

33. **No test** You toss a coin 10 times to perform a test of $H_0: p = 0.5$ that the coin is balanced against $H_a: p \neq 0.5$.

34. **No test** A college president says, "99% of the alumni support my firing of Coach Boggs." You contact an SRS of 200 of the college's 15,000 living alumni to perform a test of $H_0: p = 0.99$ versus $H_a: p < 0.99$.

35. **Home computers** Refer to Exercise 31. In Jason's SRS, 41 of the students had a computer at home. pg 556

(a) Calculate the test statistic.

(b) Find the P-value using Table A or technology. Show this result as an area under a standard Normal curve.

36. **Walking to school** Refer to Exercise 32. For DeAnna's survey, 17 students in the sample said they typically walk to school.

(a) Calculate the test statistic.

(b) Find the P-value using Table A or technology. Show this result as an area under a standard Normal curve.

37. **Significance tests** A test of $H_0: p = 0.5$ versus $H_a: p > 0.5$ has test statistic $z = 2.19$.

(a) What conclusion would you draw at the 5% significance level? At the 1% level?

(b) If the alternative hypothesis were $H_a: p \neq 0.5$, what conclusion would you draw at the 5% significance level? At the 1% level?

38. **Significance tests** A test of $H_0: p = 0.65$ against $H_a: p < 0.65$ has test statistic $z = -1.78$.

(a) What conclusion would you draw at the 5% significance level? At the 1% level?

(b) If the alternative hypothesis were $H_a: p \neq 0.65$, what conclusion would you draw at the 5% significance level? At the 1% level?

39. **Better parking** A local high school makes a change that should improve student satisfaction with the parking situation. Before the change, 37% of the school's students approved of the parking that was provided. After the change, the principal surveys an SRS of 200 of the over 2500 students at the school. In all, 83 students say that they approve of the new parking arrangement. The principal cites this as evidence that the change was effective. Perform a test of the principal's claim at the $\alpha = 0.05$ significance level.

pg 559

40. **Side effects** A drug manufacturer claims that less than 10% of patients who take its new drug for treating Alzheimer's disease will experience nausea. To test this claim, researchers conduct an experiment. They give the new drug to a random sample of 300 out of 5000 Alzheimer's patients whose families have given informed consent for the patients to participate in the study. In all, 25 of the subjects experience nausea. Use these data to perform a test of the drug manufacturer's claim at the $\alpha = 0.05$ significance level.

41. **Are boys more likely?** We hear that newborn babies are more likely to be boys than girls. Is this true? A random sample of 25,468 firstborn children included 13,173 boys.[11]

(a) Do these data give convincing evidence that firstborn children are more likely to be boys than girls?

(b) To what population can the results of this study be generalized: all children or all firstborn children? Justify your answer.

42. **Fresh coffee** People of taste are supposed to prefer fresh-brewed coffee to the instant variety. On the other hand, perhaps many coffee drinkers just want their caffeine fix. A skeptic claims that only half of all coffee drinkers prefer fresh-brewed coffee. To test this claim, we ask a random sample of 50 coffee drinkers in a small city to take part in a study. Each person tastes two unmarked cups—one containing instant coffee and one containing fresh-brewed coffee—and says which he or she prefers. We find that 36 of the 50 choose the fresh coffee.

(a) Do these results give convincing evidence that coffee drinkers favor fresh-brewed over instant coffee?

(b) We presented the two cups to each coffee drinker in a random order, so that some people tasted the fresh coffee first, while others drank the instant coffee first. Why do you think we did this?

43. **Bullies in middle school** A University of Illinois study on aggressive behavior surveyed a random sample of 558 middle school students. When asked to describe their behavior in the last 30 days, 445 students said their behavior included physical aggression, social ridicule, teasing, name-calling, and issuing threats. This behavior was not defined as bullying in the questionnaire.[12] Is this evidence that more than three-quarters of middle school students engage in bullying behavior? To find out, Maurice decides to perform a significance test. Unfortunately, he made a few errors along the way. Your job is to spot the mistakes and correct them.

$$H_0: p = 0.75$$

$$H_a: \hat{p} > 0.797$$

where p = the true mean proportion of middle school students who engaged in bullying.

- A random sample of 558 middle school students was surveyed.
- $558(0.797) = 444.73$ is at least 10.

$$z = \frac{0.75 - 0.797}{\frac{\sqrt{0.797(0.203)}}{445}} = -2.46; P\text{-value} = 2(0.0069) = 0.0138$$

The probability that the null hypothesis is true is only 0.0138, so we reject H_0. This proves that more than three-quarters of the school engaged in bullying behavior.

44. **Is this coin fair?** The French naturalist Count Buffon (1707–1788) tossed a coin 4040 times. He got 2048 heads. That's a bit more than one-half. Is this evidence that Count Buffon's coin was not balanced? To find out, Luisa decides to perform a significance test. Unfortunately, she made a few errors along the way. Your job is to spot the mistakes and correct them.

$$H_0: \mu > 0.5$$

$$H_a: \bar{x} = 0.5$$

- **10%:** $4040(0.5) = 2020$ and $4040(1 - 0.5) = 2020$ are both at least 10.
- **Large Counts:** There are at least 40,400 coins in the world.

$$t = \frac{0.5 - 0.507}{\sqrt{\frac{0.5(0.5)}{4040}}} = -0.89; P\text{-value} = 1 - 0.1867 = 0.8133$$

Reject H_0 because the P-value is so large and conclude that the coin is fair.

45. **Teen drivers** A state's Division of Motor Vehicles (DMV) claims that 60% of teens pass their driving test on the first attempt. An investigative reporter examines an SRS of the DMV records for 125 teens; 86 of them passed the test on their first try. Is there convincing evidence at the $\alpha = 0.05$ significance level that the DMV's claim is incorrect?

pg 562

46. **We want to be rich** In a recent year, 73% of first-year college students responding to a national survey identified "being very well-off financially" as an important personal goal. A state university finds that 132 of an SRS of 200 of its first-year students say that this goal is important. Is there convincing evidence at the $\alpha = 0.05$ significance level that the proportion of all first-year students at this university who think being very well-off is important differs from the national value, 73%?

47. **Teen drivers** Refer to Exercise 45.

(a) Construct and interpret a 95% confidence interval for the proportion of all teens in the state who passed their driving test on the first attempt.

(b) Explain what the interval in part (a) tells you about the DMV's claim.

48. **We want to be rich** Refer to Exercise 46.

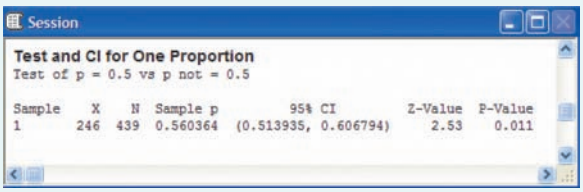
(a) Construct and interpret a 95% confidence interval for the true proportion p of all first-year students at the university who would identify being well-off as an important personal goal.

(b) Explain what the interval in part (a) tells you about whether the national value holds at this university.

49. **Do you Tweet?** In early 2012, the Pew Internet and American Life Project asked a random sample of U.S. adults, "Do you ever . . . use Twitter or another service to share updates about yourself or to see updates about others?" According to Pew, the resulting 95% confidence interval is (0.123, 0.177).[13] Does this interval provide convincing evidence that the actual proportion of U.S. adults who would say they use Twitter differs from 0.16? Justify your answer.

50. **Losing weight** A Gallup Poll found that 59% of the people in its sample said "Yes" when asked, "Would you like to lose weight?" Gallup announced: "For results based on the total sample of national adults, one can say with 95% confidence that the margin of (sampling) error is ±3 percentage points."[14] Does this interval provide convincing evidence that the actual proportion of U.S. adults who would say they want to lose weight differs from 0.55? Justify your answer.

51. **Teens and sex** The Gallup Youth Survey asked a random sample of U.S. teens aged 13 to 17 whether they thought that young people should wait to have sex until marriage.[15] The Minitab output below shows the results of a significance test and a 95% confidence interval based on the survey data.



(a) Define the parameter of interest.

(b) Check that the conditions for performing the significance test are met in this case.

(c) Interpret the P-value in context.

(d) Do these data give convincing evidence that the actual population proportion differs from 0.5? Justify your answer with appropriate evidence.

52. **Reporting cheating** What proportion of students are willing to report cheating by other students? A student project put this question to an SRS of 172 undergraduates at a large university: "You witness two students cheating on a quiz. Do you go to the professor?" The Minitab output below shows the results of a significance test and a 95% confidence interval based on the survey data.[16]

(a) Define the parameter of interest.

(b) Check that the conditions for performing the significance test are met in this case.

(c) Interpret the P-value in context.

(d) Do these data give convincing evidence that the actual population proportion differs from 0.15? Justify your answer with appropriate evidence.

53. **Better parking** Refer to Exercise 39.

(a) Describe a Type I error and a Type II error in this setting, and explain the consequences of each.

(b) The test has a power of 0.75 to detect that $p = 0.45$. Explain what this means.

(c) Identify two ways to increase the power in part (b).

54. Side effects Refer to Exercise 40.

(a) Describe a Type I error and a Type II error in this setting, and explain the consequences of each.

(b) The test has a power of 0.54 to detect that $p = 0.07$. Explain what this means.

(c) Identify two ways to increase the power in part (b).

55. Error probabilities You read that a statistical test at significance level $\alpha = 0.05$ has power 0.78. What are the probabilities of Type I and Type II errors for this test?

56. Error probabilities You read that a statistical test at the $\alpha = 0.01$ level has probability 0.14 of making a Type II error when a specific alternative is true. What is the power of the test against this alternative?

57. Power A drug manufacturer claims that fewer than 10% of patients who take its new drug for treating Alzheimer's disease will experience nausea. To test this claim, a significance test is carried out of

$$H_0: p = 0.10$$

$$H_a: p < 0.10$$

You learn that the power of this test at the 5% significance level against the alternative $p = 0.08$ is 0.29.

(a) Explain in simple language what "power = 0.29" means in this setting.

(b) You could get higher power against the same alternative with the same α by changing the number of measurements you make. Should you make more measurements or fewer to increase power? Explain.

(c) If you decide to use $\alpha = 0.01$ in place of $\alpha = 0.05$, with no other changes in the test, will the power increase or decrease? Justify your answer.

(d) If you shift your interest to the alternative $p = 0.07$ with no other changes, will the power increase or decrease? Justify your answer.

58. What is power? You manufacture and sell a liquid product whose electrical conductivity is supposed to be 5. You plan to make six measurements of the conductivity of each lot of product. If the product meets specifications, the mean of many measurements will be 5. You will therefore test

$$H_0: \mu = 5$$

$$H_a: \mu \neq 5$$

If the true conductivity is 5.1, the liquid is not suitable for its intended use. You learn that the power of your test at the 5% significance level against the alternative $\mu = 5.1$ is 0.23.

(a) Explain in simple language what "power = 0.23" means in this setting.

(b) You could get higher power against the same alternative with the same α by changing the number of measurements you make. Should you make more measurements or fewer to increase power?

(c) If you decide to use $\alpha = 0.10$ in place of $\alpha = 0.05$, with no other changes in the test, will the power increase or decrease? Justify your answer.

(d) If you shift your interest to the alternative $\mu = 5.2$, with no other changes, will the power increase or decrease? Justify your answer.

Multiple choice: Select the best answer for Exercises 59 to 62.

59. After once again losing a football game to the archrival, a college's alumni association conducted a survey to see if alumni were in favor of firing the coach. An SRS of 100 alumni from the population of all living alumni was taken, and 64 of the alumni in the sample were in favor of firing the coach. Suppose you wish to see if a majority of living alumni are in favor of firing the coach. The appropriate test statistic is

(a) $z = \dfrac{0.64 - 0.5}{\sqrt{\dfrac{0.64(0.36)}{100}}}$

(b) $t = \dfrac{0.64 - 0.5}{\sqrt{\dfrac{0.64(0.36)}{100}}}$

(c) $z = \dfrac{0.64 - 0.5}{\sqrt{\dfrac{0.5(0.5)}{100}}}$

(d) $z = \dfrac{0.64 - 0.5}{\sqrt{\dfrac{0.64(0.36)}{64}}}$

(e) $z = \dfrac{0.5 - 0.64}{\sqrt{\dfrac{0.5(0.5)}{100}}}$

60. Which of the following is *not* a condition for performing a significance test about a population proportion p?

(a) The data should come from a random sample or randomized experiment.

(b) Both np_0 and $n(1 - p_0)$ should be at least 10.

(c) If you are sampling without replacement from a finite population, then you should sample no more than 10% of the population.

(d) The population distribution should be approximately Normal, unless the sample size is large.

(e) All of the above are conditions for performing a significance test about a population proportion.

61. The z statistic for a test of $H_0: p = 0.4$ versus $H_a: p \neq 0.4$ is $z = 2.43$. This test is

(a) not significant at either $\alpha = 0.05$ or $\alpha = 0.01$.

(b) significant at $\alpha = 0.05$ but not at $\alpha = 0.01$.

(c) significant at $\alpha = 0.01$ but not at $\alpha = 0.05$.

(d) significant at both $\alpha = 0.05$ and $\alpha = 0.01$.

(e) inconclusive because we don't know the value of $\hat{p}$.

62. Which of the following 95% confidence intervals would lead us to reject $H_0: p = 0.30$ in favor of $H_a: p \neq 0.30$ at the 5% significance level?

(a) (0.19, 0.27) (c) (0.27, 0.31) (e) None of these

(b) (0.24, 0.30) (d) (0.29, 0.38)

63. **Packaging CDs** (6.2, 5.3) A manufacturer of compact discs (CDs) wants to be sure that their CDs will fit inside the plastic cases they have bought for packaging. Both the CDs and the cases are circular. According to the supplier, the plastic cases vary Normally with mean diameter $\mu = 4.2$ inches and standard deviation $\sigma = 0.05$ inches. The CD manufacturer decides to produce CDs with mean diameter $\mu = 4$ inches. Their diameters follow a Normal distribution with $\sigma = 0.1$ inches.

(a) Let X = the diameter of a randomly selected CD and Y = the diameter of a randomly selected case. Describe the shape, center, and spread of the distribution of the random variable X − Y. What is the importance of this random variable to the CD manufacturer?

(b) Compute the probability that a randomly selected CD will fit inside a randomly selected case.

(c) The production process actually runs in batches of 100 CDs. If each of these CDs is paired with a randomly chosen plastic case, find the probability that all the CDs fit in their cases.

64. **Cash to find work?** (4.2) Will cash bonuses speed the return to work of unemployed people? The Illinois Department of Employment Security designed an experiment to find out. The subjects were 10,065 people aged 20 to 54 who were filing claims for unemployment insurance. Some were offered $500 if they found a job within 11 weeks and held it for at least 4 months. Others could tell potential employers that the state would pay the employer $500 for hiring them. A control group got neither kind of bonus.[17]

(a) Describe a completely randomized design for this experiment.

(b) How will you label the subjects for random assignment? Use Table D at line 127 to choose the first 3 subjects for the first treatment.

(c) Explain the purpose of a control group in this setting.

9.3 Tests about a Population Mean

WHAT YOU WILL LEARN By the end of the section, you should be able to:

- State and check the Random, 10%, and Normal/Large Sample conditions for performing a significance test about a population mean.
- Perform a significance test about a population mean.
- Use a confidence interval to draw a conclusion for a two-sided test about a population parameter.
- Perform a significance test about a mean difference using paired data.

Confidence intervals and significance tests for a population proportion p are based on z-values from the standard Normal distribution. Inference about a population mean μ uses a t distribution with $n - 1$ degrees of freedom, except in the rare case when the population standard deviation σ is known. We learned how to construct confidence intervals for a population mean in Section 8.3. Now we'll examine the details of testing a claim about a population mean μ.

Carrying Out a Significance Test for μ

In an earlier example, a company claimed to have developed a new AAA battery that lasts longer than its regular AAA batteries. Based on years of experience, the company knows that its regular AAA batteries last for 30 hours of continuous use, on average. An SRS of 15 new batteries lasted an average of 33.9 hours with a standard deviation of 9.8 hours. Do these data give convincing evidence that the new batteries last longer on average? To find out, we perform a test of

$$H_0: \mu = 30 \text{ hours}$$
$$H_a: \mu > 30 \text{ hours}$$

where μ is the true mean lifetime of the new deluxe AAA batteries.

Conditions In Chapter 8, we introduced conditions that should be met before we construct a confidence interval for a population mean: Random, 10% when sampling without replacement, and Normal/Large Sample. These same three conditions must be verified before performing a significance test about a population mean.

As in the previous chapter, the Normal/Large Sample condition for means is

population distribution is Normal or sample size is large ($n \geq 30$)

We often don't know whether the population distribution is Normal. But if the sample size is large ($n \geq 30$), we can safely carry out a significance test. If the sample size is small, we should examine the sample data for any obvious departures from Normality, such as strong skewness and outliers.

CONDITIONS FOR PERFORMING A SIGNIFICANCE TEST ABOUT A MEAN

- **Random:** The data come from a well-designed random sample or randomized experiment.
  - **10%:** When sampling without replacement, check that $n \leq \frac{1}{10}N$.
- **Normal/Large Sample:** The population has a Normal distribution or the sample size is large ($n \geq 30$). If the population distribution has unknown shape and $n < 30$, use a graph of the sample data to assess the Normality of the population. Do not use t procedures if the graph shows strong skewness or outliers.

Here's an example that shows how to check the conditions.

EXAMPLE Better Batteries

Checking conditions

Figure 9.11 on the next page shows a dotplot, boxplot, and Normal probability plot of the battery lifetimes for an SRS of 15 batteries.

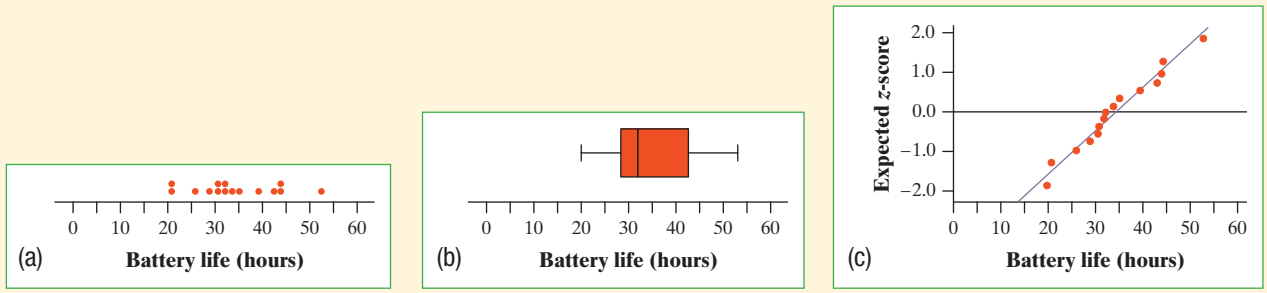


FIGURE 9.11 (a) A dotplot, (b) a boxplot, and (c) a Normal probability plot of the lifetimes of a simple random sample of 15 AAA batteries.

PROBLEM: Check the conditions for carrying out a significance test of the company's claim about its deluxe AAA batteries.

SOLUTION:

- *Random:* The company tested a simple random sample of 15 new AAA batteries.
  - 10%: Because the batteries are being sampled without replacement, we need to check that there are at least $10(15) = 150$ new AAA batteries. This seems reasonable to believe.
- *Normal/Large Sample:* We don't know if the population distribution of battery lifetimes for the company's new AAA batteries is Normal. With such a small sample size ($n = 15$), we need to graph the data to look for any departures from Normality. The dotplot and boxplot show slight right-skewness but no outliers. The Normal probability plot is fairly linear. Because none of the graphs shows any strong skewness or outliers, we should be safe performing a test about the population mean lifetime μ.

For Practice *Try Exercise* 65

Calculations: Test Statistic and *P*-Value

When performing a significance test, we do calculations assuming that the null hypothesis H_0 is true. The test statistic measures how far the sample result diverges from the parameter value specified by H_0, in standardized units. As before,

$$\text{test statistic} = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$$

For a test of $H_0: \mu = \mu_0$, our statistic is the sample mean $\bar{x}$. Its standard deviation is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

In an ideal world, our test statistic would be

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Because the population standard deviation σ is usually unknown, we use the sample standard deviation s_x in its place. The resulting test statistic has the standard error of $\bar{x}$ in the denominator

$$t = \frac{\bar{x} - \mu_0}{s_x/\sqrt{n}}$$

As we saw earlier, when the Normal/Large Sample condition is met, this statistic has approximately a t distribution with $n - 1$ degrees of freedom.

There is a small number of real-world situations in which we might know the population standard deviation σ. When this is the case, the test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

will follow a standard Normal distribution if the Normal/Large Sample condition is met. If so, then we can calculate *P*-values using Table A or technology. The TI-83/84 and TI-89's *Z*-Test option in the TESTS menu is designed for this special situation.

In Section 8.3, we used Table B to find critical values from the t distributions when constructing confidence intervals about an unknown population mean μ. Once we have calculated the test statistic, we can use Table B to find the P-value for a significance test about μ. The following example shows how this works.

EXAMPLE Better Batteries

Computing the test statistic and P-value

The battery company wants to test $H_0: \mu = 30$ versus $H_a: \mu > 30$ based on an SRS of 15 new AAA batteries with mean lifetime $\bar{x} = 33.9$ hours and standard deviation $s_x = 9.8$ hours. The test statistic is

$$\text{test statistic} = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$$

$$t = \frac{\bar{x} - \mu_0}{s_x/\sqrt{n}} = \frac{33.9 - 30}{9.8/\sqrt{15}} = 1.54$$

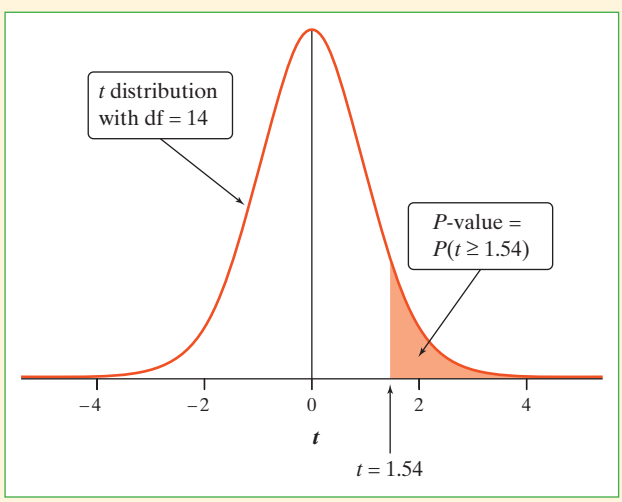


FIGURE 9.12 The P-value for a one-sided test with $t = 1.54$.

The P-value is the probability of getting a result this large or larger in the direction indicated by H_a, that is, $P(t \geq 1.54)$. Figure 9.12 shows this probability as an area under the t distribution curve with df $= 15 - 1 = 14$. We can find this P-value using Table B.

Upper-tail probability p			
df	**.10**	**.05**	**.025**
13	1.350	1.771	2.160
14	1.345	1.761	2.145
15	1.341	1.753	2.131
	80%	90%	95%
	Confidence level *C*		

Go to the df $= 14$ row. The t statistic falls between the values 1.345 and 1.761. If you look at the top of the corresponding columns in Table B, you'll find that the "Upper-tail probability p" is between 0.10 and 0.05. (See the excerpt from Table B at right.) Because we are looking for $P(t \geq 1.54)$, this is the probability we seek. That is, the P-value for this test is between 0.05 and 0.10.

As you can see, Table B gives an interval of possible P-values for a significance test. We can still draw a conclusion from the test in much the same way as if we had a single probability. In the case of the new AAA batteries, for instance, we would fail to reject $H_0: \mu = 30$ because the P-value exceeds our default $\alpha = 0.05$ significance level. We don't have convincing evidence that the company's new AAA batteries last longer than 30 hours, on average.

Table B has other limitations for finding P-values. It includes probabilities only for t distributions with degrees of freedom from 1 to 30 and then skips to df $= 40$, 50, 60, 80, 100, and 1000. (The bottom row gives probabilities for df $= \infty$, which corresponds to the standard Normal distribution.) Also, Table B shows probabilities only for positive values of t. To find a P-value for a negative value of t, we use the symmetry of the t distributions. The next example shows how we deal with both of these issues.

EXAMPLE Two-Sided Tests, Negative *t*-Values, and More

Using Table B wisely

What if you were performing a test of $H_0: \mu = 5$ versus $H_a: \mu \neq 5$ based on a sample size of $n = 37$ and obtained $t = -3.17$? Because this is a two-sided test, you are interested in the probability of getting a value of t less than or equal to -3.17 or greater than or equal to 3.17. Figure 9.13 shows the desired P-value as an area under the t distribution curve with 36 degrees of freedom. Notice that $P(t \leq -3.17) = P(t \geq 3.17)$ due to the symmetric shape of the density curve. Table B shows only positive t-values, so we will focus on $t = 3.17$.

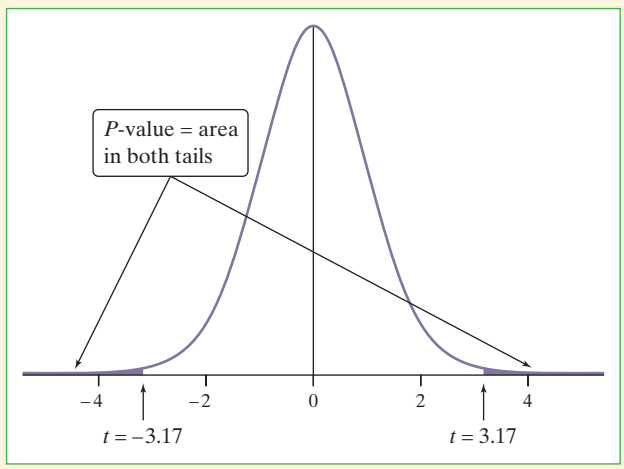


FIGURE 9.13 The P-value for a two-sided test with $t = -3.17$.

Upper-tail probability *p*			
df	**.005**	**.0025**	**.001**
29	2.756	3.038	3.396
30	2.750	3.030	3.385
40	2.704	2.971	3.307
	99%	99.5%	99.8%
	Confidence level *C*		

Because df $= 37 - 1 = 36$ is not available on the table, use df $= 30$. You might be tempted to use df $= 40$, but doing so would result in a smaller P-value than you are entitled to with df $= 36$. (In other words, you'd be cheating!) Move across the df $= 30$ row, and notice that $t = 3.17$ falls between 3.030 and 3.385. The corresponding "Upper-tail probability p" is between 0.001 and 0.0025. (See the excerpt from Table B.) For this two-sided test, the corresponding P-value would be between $2(0.001) = 0.002$ and $2(0.0025) = 0.005$.

One point from the example deserves repeating: *if the df you need isn't provided in Table B, use the next lower df that is available.* It's no fair "rounding up" to a larger df. This is like pretending that your sample size is larger than it really is. Doing so would give you a smaller P-value than is true and would make you more likely to incorrectly reject H_0 when it's true (make a Type I error).

Given the limitations of Table B, our advice is to use technology to find P-values when carrying out a significance test about a population mean.

19. TECHNOLOGY CORNER

COMPUTING *P*-VALUES FROM *t* DISTRIBUTIONS ON THE CALCULATOR

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

You can use the `tcdf` command on the TI-83/84 and TI-89 to calculate areas under a *t* distribution curve. The syntax is `tcdf(lower bound,upper bound,df)`.

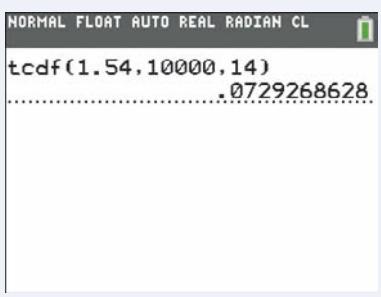
Let's use the `tcdf` command to compute the *P*-values from the last two examples.

Better batteries: To find $P(t \geq 1.54)$,

TI-83/84

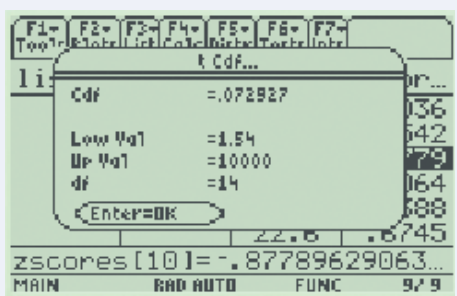
- Press 2nd VARS (DISTR) and choose `tcdf(`.

OS 2.55 or later: In the dialog box, enter these values: `lower:1.54, upper: 10000, df:14,` choose `Paste,` and then press ENTER. **Older OS:** Complete the command `tcdf(1.54,10000,14)` and press ENTER.



TI-89

- In the Stats/List Editor, press F5 (Distr) and choose `t Cdf...`.
- In the dialog box, enter these values: `Lower value:1.54, Upper value:10000, Deg of Freedom, df:14`, and then choose ENTER.



Two-sided test: To find the *P*-value for the two-sided test with df = 36 and $t = -3.17$, do `tcdf(-10000,-3.17,36)` and multiply the result by 2.

CHECK YOUR UNDERSTANDING

The makers of Aspro brand aspirin want to be sure that their tablets contain the right amount of active ingredient (acetylsalicylic acid). So they inspect a random sample of 36 tablets from a batch in production. When the production process is working properly, Aspro tablets have an average of $\mu = 320$ milligrams (mg) of active ingredient. The amount of active ingredient in the 36 selected tablets has mean 319 mg and standard deviation 3 mg.

1. State appropriate hypotheses for a significance test in this setting.
2. Check that the conditions are met for carrying out the test.
3. Calculate the test statistic. Show your work.
4. Use Table B to find the *P*-value. Then use technology to get a more accurate result. What conclusion would you draw?

The One-Sample *t* Test

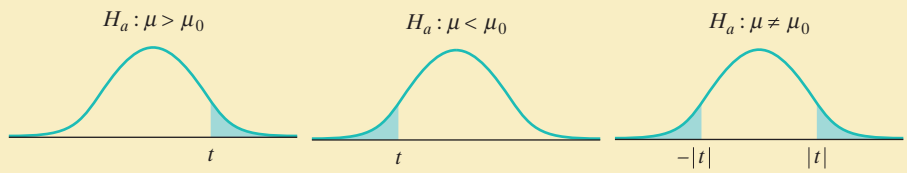
When the conditions are met, we can test a claim about a population mean μ using a **one-sample *t* test for a mean**. Here are the details.

ONE-SAMPLE *t* TEST FOR A MEAN

Suppose the conditions are met. To test the hypothesis $H_0: \mu = \mu_0$, compute the one-sample t statistic

$$t = \frac{\bar{x} - \mu_0}{s_x/\sqrt{n}}$$

Find the P-value by calculating the probability of getting a t statistic this large or larger in the direction specified by the alternative hypothesis H_a in a t distribution with df $= n - 1$:



Now we are ready to test a claim about an unknown population mean. Once again, we follow the four-step process.

EXAMPLE Healthy Streams

STEP 4

Performing a significance test about μ

The level of dissolved oxygen (DO) in a stream or river is an important indicator of the water's ability to support aquatic life. A researcher measures the DO level at 15 randomly chosen locations along a stream. Here are the results in milligrams per liter (mg/l):

4.53	5.04	3.29	5.23	4.13	5.50	4.83	4.40
5.42	6.38	4.01	4.66	2.87	5.73	5.55	

A dissolved oxygen level below 5 mg/l puts aquatic life at risk.

PROBLEM:

(a) Do we have convincing evidence at the $\alpha = 0.05$ significance level that aquatic life in this stream is at risk?

(b) Given your conclusion in part (a), which kind of mistake—a Type I error or a Type II error—could you have made? Explain what this mistake would mean in context.

SOLUTION:

(a) We will follow the four-step process.

STATE: We want to test a claim about the true mean dissolved oxygen level μ in this stream at the $\alpha = 0.05$ level. Our hypotheses are

$$H_0: \mu = 5$$
$$H_a: \mu < 5$$

AP® EXAM TIP It is not enough just to make a graph of the data on your calculator when assessing Normality. You must *sketch* the graph on your paper to receive credit. You don't have to draw multiple graphs—any appropriate graph will do.

PLAN: If the conditions are met, we should do a one-sample t test for μ.

- *Random:* The researcher measured the DO level at 15 randomly chosen locations.
  - 10%: There is an infinite number of possible locations along the stream, so it isn't necessary to check the 10% condition.
- *Normal/Large Sample:* We don't know whether the population distribution of DO levels at all points along the stream is Normal. With such a small sample size ($n = 15$), we need to graph the data to see if it's safe to use t procedures. Figure 9.14 shows our hand sketches of a calculator histogram, boxplot, and Normal probability plot for these data. The histogram looks roughly symmetric; the boxplot shows no outliers; and the Normal probability plot is fairly linear. With no outliers or strong skewness, the t procedures should be pretty accurate.

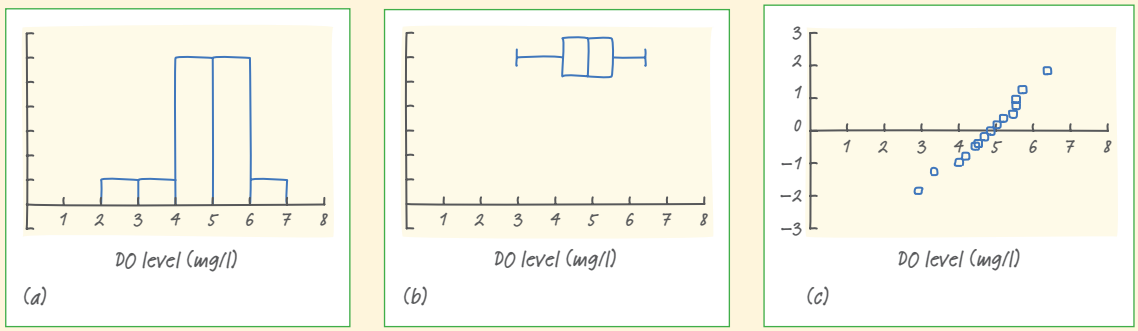


FIGURE 9.14 Sketches of (a) a histogram, (b) a boxplot, and (c) a Normal probability plot for the dissolved oxygen (DO) readings in the sample, in mg/l.

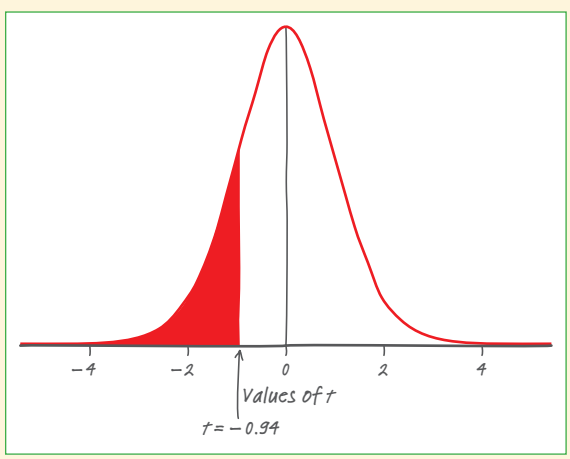
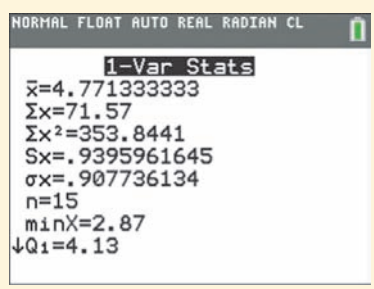


FIGURE 9.15 The P-value for a one-sided test when $t = -0.94$.

DO: We entered the data into our calculator and did 1-Var Stats (see screen shot). The sample mean is $\bar{x} = 4.771$ and the sample standard deviation is $s_x = 0.9396$.



- *Test statistic*

$$t = \frac{\bar{x} - \mu_0}{s_x/\sqrt{n}} = \frac{4.771 - 5}{0.9396/\sqrt{15}} = -0.94$$

- *P-value* The P-value is the area to the left of $t = -0.94$ under the t distribution curve with degrees of freedom $df = 15 - 1 = 14$. Figure 9.15 shows this probability.

Using the table: Table B shows only areas in the upper tail of the distribution. Because the t distributions are symmetric, $P(t \le -0.94) = P(t \ge 0.94)$. Search the $df = 14$ row of Table B for entries that bracket $t = 0.94$ (see the excerpt at right). Because the observed t lies between 0.868 and 1.076, the P-value lies between 0.15 and 0.20.

Upper-tail probability p			
df	**.25**	**.20**	**.15**
13	.694	.870	1.079
14	.692	.868	1.076
15	.691	.866	1.074
	50%	60%	70%
	Confidence level C		

Using technology: We can find the exact *P*-value using a calculator: `tcdf(lower:-100, upper:-0.94, df:14)` = 0.1816.

CONCLUDE: Because the *P*-value, 0.1816, is greater than our $\alpha = 0.05$ significance level, we fail to reject H_0. We don't have convincing evidence that the mean DO level in the stream is less than 5 mg/l.

(b) Because we decided not to reject H_0 in part (a), we could have made a Type II error (failing to reject H_0 when H_0 is false). If we did, then the mean dissolved oxygen level μ in the stream is actually less than 5 mg/l, but we didn't find convincing evidence of that with our significance test.

For Practice *Try Exercise* 73

FIGURE 9.16 Minitab output for the one-sample *t* test from the dissolved oxygen example.

Because the *t* procedures are so common, all statistical software packages will do the calculations for you. Figure 9.16 shows the output from Minitab for the one-sample *t* test in the previous example. Note that the results match!

You can also use your calculator to carry out a one-sample *t* test. But be sure to read the AP® exam tip at the end of the Technology Corner.

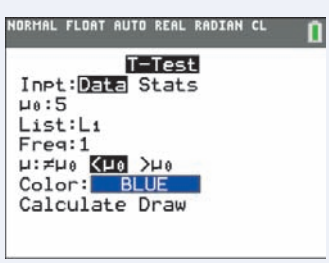
20. TECHNOLOGY CORNER ONE-SAMPLE *t* TEST FOR A MEAN ON THE CALCULATOR

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

You can perform a one-sample *t* test using either raw data or summary statistics on the TI-83/84 or TI-89. Let's use the calculator to carry out the test of H_0: $\mu = 5$ versus H_a: $\mu < 5$ from the dissolved oxygen example. Start by entering the sample data in L1/list1. Then, to do the test:

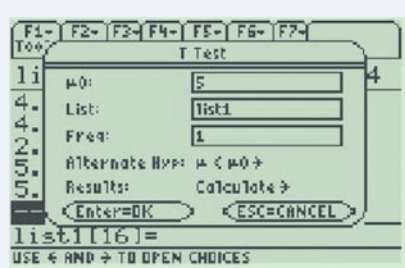
TI-83/84

- Press `STAT`, choose TESTS and `T-Test`.
- Adjust your settings as shown.

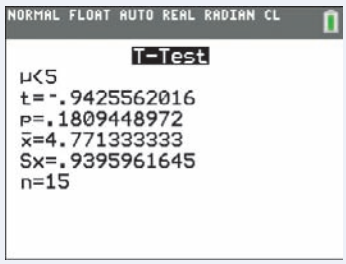


TI-89

- Press `2nd` `F1` ([F6]) and choose `T-Test`.
- Adjust your settings as shown.

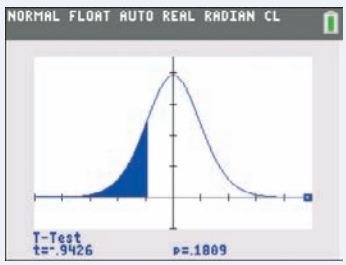


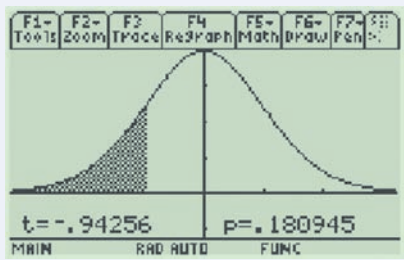
If you select "Calculate," the following screen appears:



The test statistic is $t = -0.94$ and the P-value is 0.1809.

If you specify "Draw," you see a t distribution curve (df = 14) with the lower tail shaded.





Note: If you are given summary statistics instead of the original data, you would select the option "Stats" instead of "Data" in the first screen.

AP® EXAM TIP Remember: if you just give calculator results with no work, and one or more values are wrong, you probably won't get any credit for the "Do" step. If you opt for the calculator-only method, name the procedure (*t* test) and report the test statistic ($t = -0.94$), degrees of freedom (df = 14), and *P*-value (0.1809).

CHECK YOUR UNDERSTANDING

A college professor suspects that students at his school are getting less than 8 hours of sleep a night, on average. To test his belief, the professor asks a random sample of 28 students, "How much sleep did you get last night?" Here are the data (in hours):

9 6 8 6 8 8 6 6.5 6 7 9 4 3 4 5 6 11 6 3 6 6 10 7 8 4.5 9 7 7

Do these data provide convincing evidence at the $\alpha = 0.05$ significance level in support of the professor's suspicion?

Two-Sided Tests and Confidence Intervals

Now let's look at an example involving a two-sided test.

EXAMPLE Juicy Pineapples

STEP 4

A two-sided test

At the Hawaii Pineapple Company, managers are interested in the sizes of the pineapples grown in the company's fields. Last year, the mean weight of the pineapples harvested from one large field was 31 ounces. A different irrigation system was installed in this field after the growing season. Managers wonder how this change will affect the mean weight of future pineapples grown in the field. To

find out, they select and weigh a random sample of 50 pineapples from this year's crop. The Minitab output below summarizes the data.

Descriptive Statistics: Weight (oz)

Variable	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Weight (oz)	50	31.935	0.339	2.394	26.491	29.990	31.739	34.115	35.547

PROBLEM:

(a) Do these data give convincing evidence that the mean weight of pineapples produced in the field has changed this year?

(b) Can we conclude that the different irrigation system caused a change in the mean weight of pineapples produced? Explain your answer.

SOLUTION:

(a) **STATE:** We want to perform a test of

$$H_0: \mu = 31$$
$$H_a: \mu \neq 31$$

where μ = the mean weight (in ounces) of all pineapples grown in the field this year. Because no significance level is given, we'll use $\alpha = 0.05$.

PLAN: If the conditions are met, we should conduct a one-sample t test for μ.

- *Random:* The data came from a random sample of 50 pineapples from this year's crop.
  - 10%: There need to be at least $10(50) = 500$ pineapples in the field because managers are sampling without replacement. We would expect many more than 500 pineapples in a "large field."
- *Normal/Large Sample:* We don't know whether the population distribution of pineapple weights this year is Normally distributed. But $n = 50 \geq 30$, so the large sample size makes it OK to use t procedures.

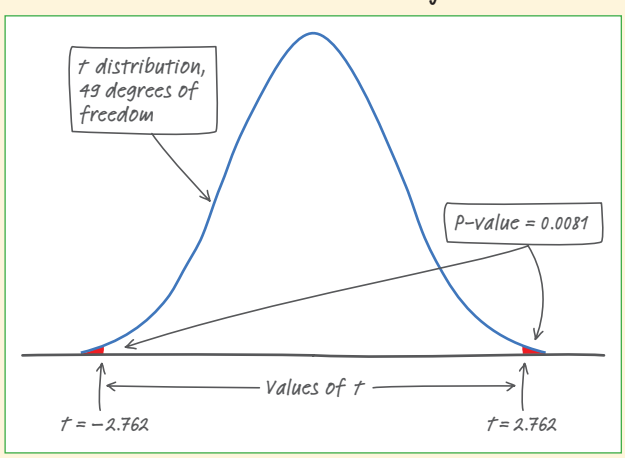


FIGURE 9.17 The P-value for the two-sided test with $t = 2.762$.

DO: From the Minitab output, $\bar{x} = 31.935$ ounces and $s_x = 2.394$ ounces.

- *Test statistic*

$$t = \frac{\bar{x} - \mu_0}{s_x/\sqrt{n}} = \frac{31.935 - 31}{2.394/\sqrt{50}} = 2.762$$

- *P-value* Figure 9.17 displays the P-value for this two-sided test as an area under the t distribution curve with $50 - 1 = 49$ degrees of freedom.

Using the table: Table B doesn't have an entry for df = 49, so we have to use the more conservative df = 40. As the excerpt below shows, the upper-tail probability is between 0.0025 and 0.005. So the desired P-value for this two-sided test is between $2(0.0025) = 0.005$ and $2(0.005) = 0.01$.

Upper-tail probability p			
df	.005	.0025	.001
30	2.750	3.030	3.385
40	2.704	2.971	3.307
50	2.678	2.937	3.261
	99%	99.5%	99.8%
	Confidence level C		

Using technology: The calculator's T-Test command gives $t = 2.762$ and P-value 0.0081 using df = 49.

CONCLUDE: Because the P-value, 0.0081, is less than $\alpha = 0.05$, we reject H_0. We have convincing evidence that the mean weight of the pineapples grown this year is not 31 ounces.

(b) No. This was not a comparative experiment, so we cannot infer causation. It is possible that other things besides the irrigation system changed from last year's growing season. Maybe the weather was different this year, and that's why the pineapples have a different mean weight than last year.

For Practice *Try Exercise* 77

The significance test in the previous example gives convincing evidence that the mean weight μ of the pineapples grown in the field this year differs from last year's 31 ounces. Unfortunately, the test doesn't give us an idea of what the actual value of μ is. For that, we need a confidence interval.

EXAMPLE Juicy Pineapples

Confidence intervals give more information

Minitab output for a significance test and confidence interval based on the pineapple data is shown below. The test statistic and *P*-value match what we got earlier (up to rounding).

One-Sample T: Weight (oz)

```
Test of mu = 31 vs not = 31
Variable      N   Mean   StDev SE Mean      95% CI         T     P
Weight (oz)  50 31.935 2.394   0.339  (31.255, 32.616) 2.76 0.008
```

The 95% confidence interval for the mean weight of all the pineapples grown in the field this year is 31.255 to 32.616 ounces. We are 95% confident that this interval captures the true mean weight μ of this year's pineapple crop.

As with proportions, there is a link between a two-sided test at significance level α and a $100(1-\alpha)\%$ confidence interval for a population mean μ. For the pineapples, the two-sided test at $\alpha = 0.05$ rejects $H_0: \mu = 31$ in favor of $H_a: \mu \neq 31$. The corresponding 95% confidence interval does not include 31 as a plausible value of the parameter μ. In other words, the test and interval lead to the same conclusion about H_0. But the confidence interval provides much more information: a set of plausible values for the population mean.

The connection between two-sided tests and confidence intervals is even stronger for means than it was for proportions. That's because both inference methods for means use the standard error of $\bar{x}$ in the calculations.

$$\text{test statistic: } t = \frac{\bar{x} - \mu_0}{s_x/\sqrt{n}} \qquad \text{Confidence interval: } \bar{x} \pm t^* \frac{s_x}{\sqrt{n}}$$

When the two-sided significance test at level α rejects $H_0: \mu = \mu_0$, the $100(1-\alpha)\%$ confidence interval for μ will not contain the hypothesized value μ_0. And when the test fails to reject the null hypothesis, the confidence interval will contain μ_0.

THINK ABOUT IT

Is there a connection between one-sided tests and confidence intervals for a population mean? As you might expect, the answer is yes. But the link is more complicated. Consider a one-sided test of $H_0: \mu = 10$ versus $H_a: \mu > 10$ based on an SRS of 30 observations. With df $= 30 - 1 = 29$, Table B says that the test will reject H_0 at $\alpha = 0.05$ if the test statistic t is greater than 1.699. For this to happen, the sample mean $\bar{x}$ would have to exceed $\mu_0 = 10$ by more than 1.699 standardized units.

Table B also shows that $t^* = 1.699$ is the critical value for a 90% confidence interval. That is, a 90% confidence interval will extend 1.699 standardized units

on either side of the sample mean $\bar{x}$. If $\bar{x}$ exceeds 10 by more than 1.699 standardized units, the resulting interval will not include 10. And the one-sided test will reject $H_0: \mu = 10$. There's the link: our one-sided test at $\alpha = 0.05$ gives the same conclusion about H_0 as a 90% confidence interval for μ.

CHECK YOUR UNDERSTANDING

The health director of a large company is concerned about the effects of stress on the company's middle-aged male employees. According to the National Center for Health Statistics, the mean systolic blood pressure for males 35 to 44 years of age is 128. The health director examines the medical records of a random sample of 72 male employees in this age group. The Minitab output displays the results of a significance test and a confidence interval.

```
Session

One-Sample T

Test of mu = 128 vs not = 128

 N    Mean  StDev  SE Mean        95% CI          T      P
72  129.93  14.90     1.76  (126.43, 133.43)  1.10  0.275
```

1. Do the results of the significance test give convincing evidence that the mean blood pressure for all the company's middle-aged male employees differs from the national average? Justify your answer.

2. Interpret the 95% confidence interval in context. Explain how the confidence interval leads to the same conclusion as in Question 1.

Inference for Means: Paired Data

Study designs that involve making two observations on the same individual, or one observation on each of two similar individuals, yield **paired data**. When paired data result from measuring the same quantitative variable twice, we can make comparisons by analyzing the differences in each pair. If the conditions for inference are met, we can use one-sample t procedures to perform inference about the mean difference μ_d. (These methods are sometimes called **paired t procedures**.) An example should help illustrate what we mean.

EXAMPLE Is Caffeine Dependence Real?

STEP 4

Paired data and one-sample t procedures

Researchers designed an experiment to study the effects of caffeine withdrawal. They recruited 11 volunteers who were diagnosed as being caffeine dependent to serve as subjects. Each subject was barred from coffee, colas, and other substances with caffeine for the duration of the experiment. During one 2-day period, subjects took capsules containing their normal caffeine intake. During another 2-day period, they took placebo capsules. The order in which subjects took caffeine and the placebo was randomized. At the end of each 2-day period, a test for depression was given to all 11 subjects. Researchers wanted to know whether being deprived of caffeine would lead to an increase in depression.[18]

The table below contains data on the subjects' scores on the depression test. Higher scores show more symptoms of depression. For each subject, we calculated the difference in test scores following each of the two treatments (placebo – caffeine). We chose this order of subtraction to get mostly positive values.

Results of a caffeine-deprivation study

Subject	Depression (caffeine)	Depression (placebo)	Difference (placebo – caffeine)
1	5	16	11
2	5	23	18
3	4	5	1
4	3	7	4
5	8	14	6
6	5	24	19
7	0	6	6
8	0	3	3
9	2	15	13
10	11	12	1
11	1	0	–1

PROBLEM:

(a) Why did researchers randomly assign the order in which subjects received placebo and caffeine?

(b) Carry out a test to investigate the researchers' question.

SOLUTION:

(a) Researchers want to be able to conclude that any statistically significant change in depression score is due to the treatments themselves and not to some other variable. One obvious concern is the order of the treatments. Suppose that caffeine were given to all the subjects during the first 2-day period. What if the weather were nicer on these 2 days than during the second 2-day period when all subjects were given a placebo? Researchers wouldn't be able to tell if a large increase in the mean depression score is due to the difference in weather or due to the treatments. Random assignment of the caffeine and placebo to the two time periods in the experiment should help ensure that no other variable (like the weather) is systematically affecting subjects' responses.

(b) We'll follow the four-step process.

STATE: If caffeine deprivation has no effect on depression, then we would expect the actual mean difference in depression scores to be 0. We therefore want to test the hypotheses

$$H_0: \mu_d = 0$$
$$H_a: \mu_d > 0$$

where μ_d is the true mean difference (placebo – caffeine) in depression score for subjects like these. Because no significance level is given, we'll use $\alpha = 0.05$.

It is uncommon for the subjects in an experiment to be randomly selected from some larger population. In fact, most experiments use recruited volunteers as subjects. When there is no sampling, we don't need to check the 10% condition.

PLAN: If the conditions are met, we should conduct a paired t test for μ_d.

- *Random:* Researchers randomly assigned the treatments—placebo then caffeine, caffeine then placebo—to the subjects.
  - 10%: We aren't sampling, so it isn't necessary to check the 10% condition.

• *Normal/Large Sample:* We don't know whether the actual distribution of difference in depression scores (placebo − caffeine) for subjects like these is Normal. With such a small sample size ($n = 11$), we need to graph the data to see if it's safe to use t procedures. Figure 9.18 shows hand sketches of a calculator histogram, boxplot, and Normal probability plot for these data. The histogram has an irregular shape with so few values; the boxplot shows some right skewness but no outliers; and the Normal probability plot is slightly curved, indicating mild skewness. With no outliers or strong skewness, the t procedures should be fairly accurate.

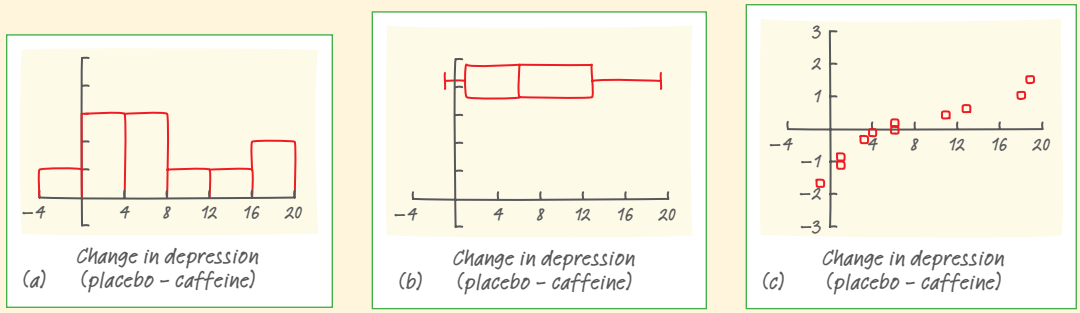


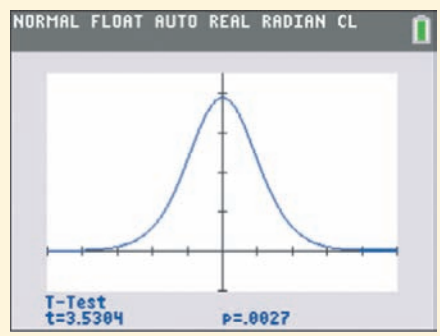
FIGURE 9.18 Sketches of (a) a histogram, (b) a boxplot, and (c) a Normal probability plot of the change in depression scores (placebo − caffeine) for the 11 subjects in the caffeine experiment.

DO: We entered the differences in list1 and then used the calculator's t test command with the "Draw" option.

• *Test statistic* $t = 3.53$

• *P-value* 0.0027, which is the area to the right of $t = 3.53$ on the t distribution curve with $df = 11 - 1 = 10$.

Note: The calculator doesn't report the degrees of freedom, but you should.



Just by looking at the data, it appears that the true mean change in depression score μ_d is greater than 0. However, it's possible that there has been no change and we got a result this much larger than $\mu_d = 0$ by the luck of the random assignment. The significance test tells us whether this explanation is plausible.

CONCLUDE: Because the P-value of 0.0027 is less than $\alpha = 0.05$, we reject $H_0: \mu_d = 0$. We have convincing evidence that the true mean difference (placebo − caffeine) in depression score is positive for subjects like these.

For Practice *Try Exercise* 85

A few follow-up comments about this example are in order.

1. We could have calculated the test statistic in the example using the formula

$$t = \frac{\bar{x}_d - \mu_0}{s_d/\sqrt{n}} = \frac{7.364 - 0}{6.918/\sqrt{11}} = 3.53$$

and obtained the P-value using Table B or technology. Check with your teacher on whether the calculator-only method is acceptable. *Be sure to report the degrees of freedom with any t procedure, even if technology doesn't.*

2. The subjects in this experiment were *not* chosen at random from the population of caffeine-dependent individuals. As a result, we can't generalize our findings to *all* caffeine-dependent people—only to people like the ones who took part in this experiment.

3. Because researchers randomly assigned the treatments, they can make an inference about cause and effect. The data from this experiment provide convincing evidence that depriving caffeine-dependent subjects like these of caffeine causes an average increase in depression scores.

Until now, we have only used one-sample t procedures in settings involving random sampling. The paired data in the caffeine example came from a matched pairs experiment, in which each subject received both treatments in a random order. A coin toss or some other chance process was used to carry out the random assignment. Why is it legitimate to use a t distribution to perform inference about the parameter μ in a randomized experiment? The answer to that question will have to wait until the next chapter.

CHECK YOUR UNDERSTANDING

Refer to the Data Exploration from Chapter 4 on page 257. Do the data give convincing evidence at the $\alpha = 0.05$ significance level that filling tires with nitrogen instead of air decreases pressure loss?

Using Tests Wisely

Significance tests are widely used in reporting the results of research in many fields. New drugs require significant evidence of effectiveness and safety. Courts ask about statistical significance in hearing discrimination cases. Marketers want to know whether a new ad campaign significantly outperforms the old one, and medical researchers want to know whether a new therapy performs significantly better. In all these uses, statistical significance is valued because it points to an effect that is unlikely to occur simply by chance.

Carrying out a significance test is often quite simple, especially if you use technology. Using tests wisely is not so simple. Here are some points to keep in mind when using or interpreting significance tests.

Determining Sample Size How large a sample should researchers take when they plan to carry out a significance test? The answer depends on three factors: significance level, effect size, and the desired power of the test. Here are the questions that researchers must answer to decide how many observations they need:

1. *Significance level.* How much risk of a Type I error—rejecting the null hypothesis when H_0 is actually true—are we willing to accept? If a Type I error has serious consequences, we might opt for $\alpha = 0.01$. Otherwise, we should choose $\alpha = 0.05$ or $\alpha = 0.10$. Recall that using a higher significance level would decrease the Type II error probability and increase the power.

2. *Effect size.* How large a difference between the null parameter value and the actual parameter value is important for us to detect?
3. *Power.* What chance do we want our study to have to detect a difference of the size we think is important?

Let's illustrate typical answers to these questions using an example.

EXAMPLE Developing Stronger Bones

Planning a study

Can a 6-month exercise program increase the total body bone mineral content (TBBMC) of young women? A team of researchers is planning a study to examine this question. The researchers would like to perform a test of

$$H_0: \mu = 0$$
$$H_a: \mu > 0$$

where μ is the true mean percent change in TBBMC due to the exercise program. To decide how many subjects they should include in their study, researchers begin by answering the three questions above.

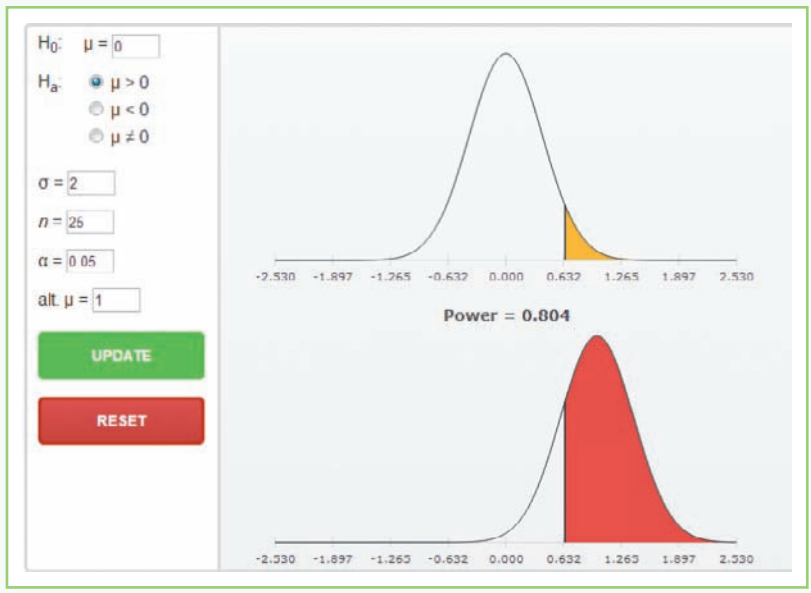
1. *Significance level.* The researchers decide that $\alpha = 0.05$ gives enough protection against declaring that the exercise program increases bone mineral content when it really doesn't (a Type I error).
2. *Effect size.* A mean increase in TBBMC of 1% would be considered important to detect.
3. *Power.* The researchers want probability at least 0.9 that a test at the chosen significance level will reject the null hypothesis $H_0: \mu = 0$ when the truth is $\mu = 1$.

The following Activity gives you a chance to investigate the sample size needed to achieve a power of 0.9 in the bone mineral content study.

ACTIVITY Investigating Power

MATERIALS:
Computer with Internet connection and display capability

In this Activity, you will use the *Statistical Power* applet at the book's Web site to determine the sample size needed for the exercise study of the previous example. Based on the results of a previous study, researchers are willing to assume that $\sigma = 2$ for the percent change in TBBMC over the 6-month period. We'll start by seeing whether or not 25 subjects are enough.



1. Go to www.whfreeman.com/tps5e and launch the *Statistical Power* applet. Enter the values: $H_0: \mu = 0$, $H_a: \mu > 0$, $\sigma = 2$, $n = 25$, $\alpha = 0.05$, and alternate $\mu = 1$. Then click "Update." What is the power? As a class, discuss what this number means in simple terms.
2. Change the significance level to 0.01. What effect does this have on the power of the test to detect $\mu = 1$? Why?
3. The researchers decide that they are willing to risk a 5% chance of making a Type I error. Change the significance level back to $\alpha = 0.05$. Now increase the sample size to 30. What happens to the power? Why?
4. Keep increasing the sample size until the power is at least 0.90. What minimum sample size should the researchers use for their study?
5. Would the researchers need a smaller or a larger sample size to detect a mean increase of 1.5% in TBBMC? A 0.85% increase? Use the applet to investigate.
6. Summarize what you have learned about how significance level, effect size, and power influence the sample size needed for a significance test.

Here is a summary of influences on "How large a sample do I need?" from the Activity.

- If you insist on a smaller significance level (such as 1% rather than 5%), you have to take a larger sample. A smaller significance level requires stronger evidence to reject the null hypothesis.
- If you insist on higher power (such as 0.99 rather than 0.90), you will need a larger sample. Higher power gives a better chance of detecting a difference when it really exists.
- At any significance level and desired power, detecting a small difference between the null and alternative parameter values requires a larger sample than detecting a large difference.

Statistical Significance and Practical Importance When a null hypothesis ("no effect" or "no difference") can be rejected at the usual levels ($\alpha = 0.05$ or $\alpha = 0.01$), there is convincing evidence of a difference. But that difference may be very small. When large samples are available, even tiny deviations from the null hypothesis will be significant.

EXAMPLE Wound Healing Time

Significant doesn't mean important

Suppose we're testing a new antibacterial cream, "Formulation NS," on a small cut made on the inner forearm. We know from previous research that with no medication, the mean healing time (defined as the time for the scab to fall off) is 7.6 days. The claim we want to test here is that Formulation NS speeds healing. We will use a 5% significance level.

Procedure: We cut a random sample of 250 college students and apply Formulation NS to the wounds. The mean healing time for these subjects is $\bar{x} = 7.5$ days and the standard deviation is $s_x = 0.9$ days.

Discussion: We want to test a claim about the mean healing time μ in the population of college students whose cuts are treated with Formulation NS. Our hypotheses are

$$H_0: \mu = 7.6 \text{ days}$$
$$H_a: \mu < 7.6 \text{ days}$$

An examination of the data reveals no outliers or strong skewness, so the conditions for performing a one-sample t test are met. We carry out the test and find that $t = -1.76$ and P-value $= 0.04$ with df $= 249$. Because 0.04 is less than $\alpha = 0.05$, we reject H_0. We have convincing evidence that Formulation NS reduces the average healing time. However, this result is not practically important. Having your scab fall off one-tenth a day sooner is no big deal.

Remember the wise saying: *Statistical significance is not the same thing as practical importance*. The remedy for attaching too much importance to statistical significance is to pay attention to the actual data as well as to the P-value. Plot your data and examine them carefully. Are there outliers or other departures from a consistent pattern? A few outlying observations can produce highly significant results if you blindly apply common significance tests. Outliers can also destroy the significance of otherwise-convincing data.

The foolish user of statistics who feeds the data to a calculator or computer without exploratory analysis will often be embarrassed. Is the difference you are seeking visible in your plots? If not, ask yourself whether the difference is large enough to be practically important. Give a confidence interval for the parameter in which you are interested. A confidence interval gives a set of plausible values for the parameter rather than simply asking if the observed result is too surprising to occur by chance alone when H_0 is true. Confidence intervals are not used as often as they should be, whereas significance tests are perhaps overused.

Beware of Multiple Analyses Statistical significance ought to mean that you have found a difference that you were looking for. The reasoning behind statistical significance works well if you decide what difference you are seeking, design a study to search for it, and use a significance test to weigh the evidence you get. In other settings, significance may have little meaning.

EXAMPLE Cell Phones and Brain Cancer

Don't search for significance!

Might the radiation from cell phones be harmful to users? Many studies have found little or no connection between using cell phones and various illnesses. Here is part of a news account of one study:

> A hospital study that compared brain cancer patients and a similar group without brain cancer found no statistically significant difference between brain cancer rates for the two groups. But when 20 distinct types of brain cancer were considered separately, a significant difference in brain cancer rates was found for one rare type. Puzzlingly, however, this risk appeared to decrease rather than increase with greater mobile phone use.[19]

Think for a moment. Suppose that the 20 null hypotheses for these 20 significance tests are all true. Then each test has a 5% chance of being significant at the 5% level. That's what $\alpha = 0.05$ means: results this extreme occur only 5% of the time just by chance when the null hypothesis is true. We expect about 1 of 20 tests to give a significant result just by chance. Running one test and reaching the $\alpha = 0.05$ level is reasonably good evidence that you have found something; running 20 tests and reaching that level only once is not.

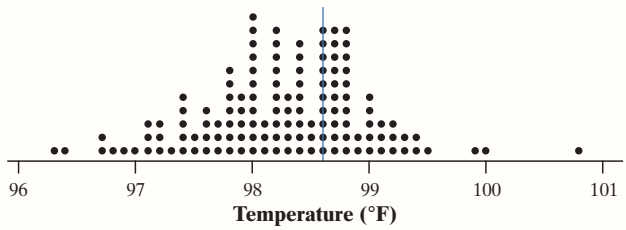
For more on the pitfalls of multiple analyses, do an Internet search for the XKCD comic about jelly beans causing acne.

Searching data for patterns is certainly legitimate. Exploratory data analysis is an important part of statistics. But the reasoning of formal inference does not apply when your search for a striking effect in the data is successful. The remedy is clear. Once you have a hypothesis, design a study to search specifically for the effect you now think is there. If the result of this study is statistically significant, you have real evidence.

case closed

Do You Have a Fever?

At the beginning of the chapter, we described a study investigating whether "normal" human body temperature is really 98.6°F. Here is a summary of the details we provided in the chapter-opening Case Study (page 537).



- The mean temperature was $\bar{x} = 98.25°F$.
- The standard deviation of the temperature readings was $s_x = 0.73°F$.
- 62.3% of the temperature readings were less than 98.6°F.

1. If "normal" body temperature really is 98.6°F, we would expect that about half of all healthy 18- to 40-year-olds will have a body temperature less than 98.6°F. Do the data from this study provide convincing evidence at the $\alpha = 0.05$ level that this is not the case?
2. The test in Question 1 has power 0.66 to detect that the actual population proportion is 0.60. Describe two changes that could be made to increase the power of the test.

Do the data provide convincing evidence that average normal body temperature is *not* 98.6°F? The computer output below shows the results of a one-sample t test and a 95% confidence interval for the population mean μ.

One-Sample T

Test of mu = 98.6 vs not = 98.6

N	Mean	StDev	SE Mean	95% CI	T	P
130	98.2500	0.7300	0.0640	(98.1233, 98.3767)	−5.47	0.000

3. What conditions must be satisfied for a one-sample t test to give valid results? Show that these conditions are met in this setting.
4. Explain how the P-value and the confidence interval lead to the same conclusion for the significance test.
5. Based on the conclusion in Question 4, which type of error could have been made: a Type I error or a Type II error? Justify your answer.

Section 9.3 Summary

- The conditions for performing a significance test of H_0: $\mu = \mu_0$ are:
  - **Random:** The data were produced by a well-designed random sample or randomized experiment.
    - **10%:** When sampling without replacement, check that the population is at least 10 times as large as the sample.

- **Normal/Large Sample:** The population distribution is Normal *or* the sample size is large ($n \geq 30$). When the sample size is small ($n < 30$), examine a graph of the sample data for any possible departures from Normality in the population. You should be safe using a t distribution as long as there is no strong skewness and no outliers are present.

- The **one-sample t test for a mean** uses the test statistic

$$t = \frac{\bar{x} - \mu_0}{s_x/\sqrt{n}}$$

with P-values calculated from the t distribution with $n - 1$ degrees of freedom.

- Confidence intervals provide additional information that significance tests do not—namely, a set of plausible values for the parameter μ. A two-sided test of $H_0: \mu = \mu_0$ at significance level α gives the same conclusion as a $100(1 - \alpha)\%$ confidence interval for μ.
- Analyze **paired data** by first taking the difference within each pair to produce a single sample. Then use one-sample t procedures.
- There are three factors that influence the sample size required for a statistical test: significance level, effect size, and the desired power of the test.
- Very small differences can be highly significant (small P-value) when a test is based on a large sample. A statistically significant difference need not be practically important.
- Many tests run at once will probably produce some significant results by chance alone, even if all the null hypotheses are true.

9.3 TECHNOLOGY CORNERS

TI-Nspire Instructions in Appendix B; HP Prime instructions on the book's Web site.

Section 9.3 Exercises

65. **Attitudes** The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures students' attitudes toward school and study habits. Scores range from 0 to 200. Higher scores indicate more positive attitudes. The mean score for U.S. college students is about 115. A teacher suspects that older students have better attitudes toward school. She gives the SSHA to an SRS of 45 of the over 1000 students at her college who are at least 30 years of age. Check the conditions for carrying out a significance test of the teacher's suspicion.

pg 575

66. **Anemia** Hemoglobin is a protein in red blood cells that carries oxygen from the lungs to body tissues. People with fewer than 12 grams of hemoglobin per deciliter of blood (g/dl) are anemic. A public health official in Jordan suspects that Jordanian children are at risk of anemia. He measures a random sample of

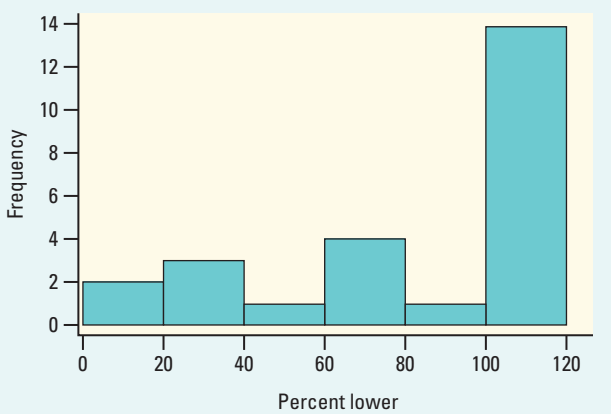
50 children. Check the conditions for carrying out a significance test of the official's suspicion.

67. **Ancient air** The composition of the earth's atmosphere may have changed over time. To try to discover the nature of the atmosphere long ago, we can examine the gas in bubbles inside ancient amber. Amber is tree resin that has hardened and been trapped in rocks. The gas in bubbles within amber should be a sample of the atmosphere at the time the amber was formed. Measurements on 9 specimens of amber from the late Cretaceous era (75 to 95 million years ago) give these percents of nitrogen:[20]

63.4 65.0 64.4 63.3 54.8 64.5 60.8 49.1 51.0

Explain why we should not carry out a one-sample t test in this setting.

68. **Paying high prices?** A retailer entered into an exclusive agreement with a supplier who guaranteed to provide all products at competitive prices. The retailer eventually began to purchase supplies from other vendors who offered better prices. The original supplier filed a lawsuit claiming violation of the agreement. In defense, the retailer had an audit performed on a random sample of 25 invoices. For each audited invoice, all purchases made from other suppliers were examined and compared with those offered by the original supplier. The percent of purchases on each invoice for which an alternative supplier offered a lower price than the original supplier was recorded.[21] For example, a data value of 38 means that the price would be lower with a different supplier for 38% of the items on the invoice. A histogram and some computer output for these data are shown below. Explain why we should not carry out a one-sample t test in this setting.



Summary statistics

Column	n	Mean	Std. Dev.	Std. Err.	Median	Min	Max	Q1	Q3
pctlower	25	77.76	32.6768	6.5353603	100	0	100	68	100

69. **Attitudes** In the study of older students' attitudes from Exercise 65, the sample mean SSHA score was 125.7 and the sample standard deviation was 29.8.

(a) Calculate the test statistic.

(b) Find the P-value using Table B. Then obtain a more precise P-value from your calculator.

70. **Anemia** For the study of Jordanian children in Exercise 66, the sample mean hemoglobin level was 11.3 mg/dl and the sample standard deviation was 1.6 mg/dl.

(a) Calculate the test statistic.

(b) Find the P-value using Table B. Then obtain a more precise P-value from your calculator.

71. **One-sided test** Suppose you carry out a significance test of $H_0: \mu = 5$ versus $H_a: \mu < 5$ based on a sample of size $n = 20$ and obtain $t = -1.81$.

(a) Find the P-value for this test using Table B or technology. What conclusion would you draw at the 5% significance level? At the 1% significance level?

(b) Redo part (a) using an alternative hypothesis of $H_a: \mu \neq 5$.

72. **Two-sided test** The one-sample t statistic from a sample of $n = 25$ observations for the two-sided test of

$$H_0: \mu = 64$$
$$H_a: \mu \neq 64$$

has the value $t = -1.12$.

(a) Find the P-value for this test using Table B or technology. What conclusion would you draw at the 5% significance level? At the 1% significance level?

(b) Redo part (a) using an alternative hypothesis of $H_a: \mu < 64$.

73. **Construction zones** Every road has one at some point—construction zones that have much lower speed limits. To see if drivers obey these lower speed limits, a police officer uses a radar gun to measure the speed (in miles per hours, or mph) of a random sample of 10 drivers in a 25 mph construction zone. Here are the data:

pg 580

27 33 32 21 30 30 29 25 27 34

(a) Is there convincing evidence that the average speed of drivers in this construction zone is greater than the posted speed limit?

(b) Given your conclusion in part (a), which kind of mistake—a Type I error or a Type II error—could you have made? Explain what this mistake would mean in context.

74. **Heat through the glass** How well materials conduct heat matters when designing houses, for example. Conductivity is measured in terms of watts of heat

power transmitted per square meter of surface per degree Celsius of temperature difference on the two sides of the material. In these units, glass has conductivity about 1. The National Institute of Standards and Technology provides exact data on properties of materials. Here are measurements of the heat conductivity of 11 randomly selected pieces of a particular type of glass:[22]

1.11 1.07 1.11 1.07 1.12 1.08 1.08 1.18 1.18 1.18 1.12

(a) Is there convincing evidence that the mean conductivity of this type of glass is greater than 1?

(b) Given your conclusion in part (a), which kind of mistake—a Type I error or a Type II error—could you have made? Explain what this mistake would mean in context.

75. **Healthy bones** The recommended daily allowance (RDA) of calcium for women between the ages of 18 and 24 years is 1200 milligrams (mg). Researchers who were involved in a large-scale study of women's bone health suspected that their participants had significantly lower calcium intakes than the RDA. To test this suspicion, the researchers measured the daily calcium intake of a random sample of 36 women from the study who fell in the desired age range. The Minitab output below displays the results of a significance test.

One-Sample T: Calcium intake (mg)

Test of mu = 1200 vs < 1200

Variable	N	Mean	StDev	SE Mean	T	P
Calcium	36	856.2	306.7	51.1	−6.73	0.000

(a) Do these data give convincing evidence to support the researchers' suspicion? Justify your answer.

(b) Interpret the *P*-value in context.

76. **Taking stock** An investor with a stock portfolio worth several hundred thousand dollars sued his broker due to the low returns he got from the portfolio at a time when the stock market did well overall. The investor's lawyer wants to compare the broker's performance against the market as a whole. He collects data on the broker's returns for a random sample of 36 weeks. Over the 10-year period that the broker has managed portfolios, stocks in the Standard & Poor's 500 index gained an average of 0.95% per week. The Minitab output below displays the results of a significance test.

One-Sample T: Return (percent)

Test of mu = 0.95 vs < 0.95

Variable	N	Mean	StDev	SE Mean	T	P
Return (percent)	36	−1.441	4.810	0.802	−2.98	0.003

(a) Do these data give convincing evidence to support the lawyer's case? Justify your answer.

(b) Interpret the *P*-value in context.

77. **Pressing pills** A drug manufacturer forms tablets by compressing a granular material that contains the active ingredient and various fillers. The hardness of a sample from each batch of tablets produced is measured to control the compression process. The target value for the hardness is $\mu = 11.5$. The hardness data for a random sample of 20 tablets are

pg 583

11.627	11.613	11.493	11.602	11.360
11.374	11.592	11.458	11.552	11.463
11.383	11.715	11.485	11.509	11.429
11.477	11.570	11.623	11.472	11.531

Is there convincing evidence at the 5% level that the mean hardness of the tablets differs from the target value?

78. **Filling cola bottles** Bottles of a popular cola are supposed to contain 300 milliliters (ml) of cola. There is some variation from bottle to bottle because the filling machinery is not perfectly precise. An inspector measures the contents of six randomly selected bottles from a single day's production. The results are

299.4 297.7 301.0 298.9 300.2 297.0

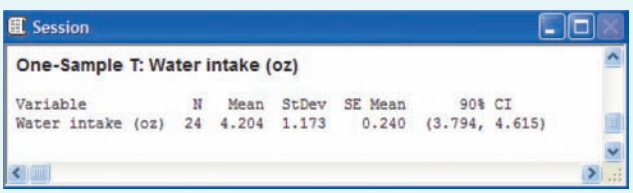
Do these data provide convincing evidence that the mean amount of cola in all the bottles filled that day differs from the target value of 300 ml?

79. **Pressing pills** Refer to Exercise 77. Construct and interpret a 95% confidence interval for the population mean μ. What additional information does the confidence interval provide?

80. **Filling cola bottles** Refer to Exercise 78. Construct and interpret a 95% confidence interval for the population mean μ. What additional information does the confidence interval provide?

81. **Fast connection?** How long does it take for a chunk of information to travel from one server to another and back on the Internet? According to the site internettrafficreport.com, a typical response time is 200 milliseconds (about one-fifth of a second). Researchers collected data on response times of a random sample of 14 servers in Europe. A graph of the data reveals no strong skewness or outliers. The following figure displays Minitab output for a one-sample *t* interval for the population mean. Is there convincing evidence at the 5% significance level that the site's claim is incorrect? Justify your answer.

82. **Water!** A blogger claims that U.S. adults drink an average of five 8-ounce glasses of water per day. Skeptical researchers ask a random sample of 24 U.S. adults about their daily water intake. A graph of the data shows a roughly symmetric shape with no outliers. The figure below displays Minitab output for a one-sample t interval for the population mean. Is there convincing evidence at the 10% significance level that the blogger's claim is incorrect? Justify your answer.



83. **Tests and CIs** The P-value for a two-sided test of the null hypothesis $H_0: \mu = 10$ is 0.06.

(a) Does the 95% confidence interval for μ include 10? Why or why not?

(b) Does the 90% confidence interval for μ include 10? Why or why not?

84. **Tests and CIs** The P-value for a two-sided test of the null hypothesis $H_0: \mu = 15$ is 0.03.

(a) Does the 99% confidence interval for μ include 15? Why or why not?

(b) Does the 95% confidence interval for μ include 15? Why or why not?

85. **Right versus left** The design of controls and instruments affects how easily people can use them. A student project investigated this effect by asking 25 right-handed students to turn a knob (with their right hands) that moved an indicator. There were two identical instruments, one with a right-hand thread (the knob turns clockwise) and the other with a left-hand thread (the knob must be turned counterclockwise). Each of the 25 students used both instruments in a random order. The following table gives the times in seconds each subject took to move the indicator a fixed distance.[23] Note that smaller times are better.

pg 586

Subject	Right thread	Left thread
1	113	137
2	105	105
3	130	133
4	101	108
5	138	115
6	118	170
7	87	103
8	116	145
9	75	78
10	96	107
11	122	84
12	103	148
13	116	147
14	107	87
15	118	166
16	103	146
17	111	123
18	104	135
19	111	112
20	89	93
21	78	76
22	100	116
23	89	78
24	85	101
25	88	123

(a) Explain why it was important to randomly assign the order in which each subject used the two knobs.

(b) The project designers hoped to show that right-handed people find right-hand threads easier to use, on average. Carry out a test at the 5% significance level to investigate this claim.

86. **Floral scents and learning** We hear that listening to Mozart improves students' performance on tests. Maybe pleasant odors have a similar effect. To test this idea, 21 subjects worked two different but roughly equivalent paper-and-pencil mazes while wearing a mask. The mask was either unscented or carried a floral scent. Each subject used both masks, in a random order. The table below gives the subjects' times (in seconds) with both masks.[24] Note that smaller times are better.

Subject	Unscented	Scented
1	30.60	37.97
2	48.43	51.57
3	60.77	56.67
4	36.07	40.47

Subject	Unscented	Scented
5	68.47	49.00
6	32.43	43.23
7	43.70	44.57
8	37.10	28.40
9	31.17	28.23
10	51.23	68.47
11	65.40	51.10
12	58.93	83.50
13	54.47	38.30
14	43.53	51.37
15	37.93	29.33
16	43.50	54.27
17	87.70	62.73
18	53.53	58.00
19	64.30	52.40
20	47.37	53.63
21	53.67	47.00

(a) Explain why it was important to randomly assign the order in which each subject used the two masks.

(b) Do these data provide convincing evidence that the floral scent improved performance, on average?

87. **Growing tomatoes** Researchers suspect that Variety A tomato plants have a higher average yield than Variety B tomato plants. To find out, researchers randomly select 10 Variety A and 10 Variety B tomato plants. Then the researchers divide in half each of 10 small plots of land in different locations. For each plot, a coin toss determines which half of the plot gets a Variety A plant; a Variety B plant goes in the other half. After harvest, they compare the yield in pounds for the plants at each location. The 10 differences in yield (Variety A − Variety B) are recorded. A graph of the differences looks roughly symmetric and single-peaked with no outliers. A paired t test on the differences yields $t = 1.295$ and P-value $= 0.1138$.

(a) State appropriate hypotheses for the paired t test. Be sure to define your parameter.

(b) What are the degrees of freedom for the paired t test?

(c) Interpret the P-value in context. What conclusion should the researchers draw?

(d) Describe a Type I error and a Type II error in this setting. Which mistake could researchers have made based on your answer to part (c)?

88. **Music and memory** Does listening to music while studying hinder students' learning? Two AP® Statistics students designed an experiment to find out. They selected a random sample of 30 students from their medium-sized high school to participate. Each subject was given 10 minutes to memorize two different lists of 20 words, once while listening to music and once in silence. The order of the two word lists was determined at random; so was the order of the treatments. The difference in the number of words recalled (music − silence) was recorded for each subject. A paired t test on the differences yielded $t = -3.01$ and P-value $= 0.0027$.

(a) State appropriate hypotheses for the paired t test. Be sure to define your parameter.

(b) What are the degrees of freedom for the paired t test?

(c) Interpret the P-value in context. What conclusion should the students draw?

(d) Describe a Type I error and a Type II error in this setting. Which mistake could students have made based on your answer to part (c)?

89. **The power of tomatoes** Refer to Exercise 87. Explain two ways that the researchers could have increased the power of the test to detect $\mu = 0.5$.

90. **Music and memory** Refer to Exercise 88. Which of the following changes would give the test a higher power to detect $\mu = -1$: using $\alpha = 0.01$ or $\alpha = 0.10$? Explain.

91. **Significance and sample size** A study with 5000 subjects reported a result that was statistically significant at the 5% level. Explain why this result might not be particularly large or important.

92. **Sampling shoppers** A marketing consultant observes 50 consecutive shoppers at a supermarket, recording how much each shopper spends in the store. Explain why it would not be wise to use these data to carry out a significance test about the mean amount spent by all shoppers at this supermarket.

93. **Do you have ESP?** A researcher looking for evidence of extrasensory perception (ESP) tests 500 subjects. Four of these subjects do significantly better ($P < 0.01$) than random guessing.

(a) Is it proper to conclude that these four people have ESP? Explain your answer.

(b) What should the researcher now do to test whether any of these four subjects have ESP?

94. **Ages of presidents** Joe is writing a report on the backgrounds of American presidents. He looks up the ages of all the presidents when they entered office. Because Joe took a statistics course, he uses these numbers to perform a significance test about the mean age of all U.S. presidents. Explain why this makes no sense.

Multiple choice: Select the best answer for Exercises 95 to 102.

95. The reason we use t procedures instead of z procedures when carrying out a test about a population mean is that

(a) z requires that the sample size be large.

(b) z requires that you know the population standard deviation σ.

(c) z requires that the data come from a random sample or randomized experiment.

(d) z requires that the population distribution be perfectly Normal.

(e) z can only be used for proportions.

96. You are testing $H_0: \mu = 10$ against $H_a: \mu < 10$ based on an SRS of 20 observations from a Normal population. The t statistic is $t = -2.25$. The P-value

(a) falls between 0.01 and 0.02.

(b) falls between 0.02 and 0.04.

(c) falls between 0.04 and 0.05.

(d) falls between 0.05 and 0.25.

(e) is greater than 0.25.

97. You are testing $H_0: \mu = 10$ against $H_a: \mu \neq 10$ based on an SRS of 15 observations from a Normal population. What values of the t statistic are statistically significant at the $\alpha = 0.005$ level?

(a) $t > 3.326$ (d) $t < -3.326$ or $t > 3.326$

(b) $t > 3.286$ (e) $t < -3.286$ or $t > 3.286$

(c) $t > 2.977$

98. After checking that conditions are met, you perform a significance test of $H_0: \mu = 1$ versus $H_a: \mu \neq 1$. You obtain a P-value of 0.022. Which of the following must be true?

(a) A 95% confidence interval for μ will include the value 1.

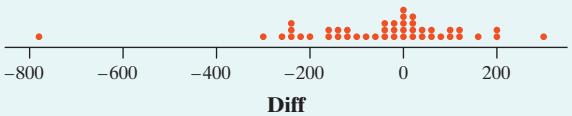
(b) A 95% confidence interval for μ will include the value 0.

(c) A 99% confidence interval for μ will include the value 1.

(d) A 99% confidence interval for μ will include the value 0.

(e) None of these is necessarily true.

99. Does Friday the 13th have an effect on people's behavior? Researchers collected data on the number of shoppers at a sample of 45 nearby grocery stores on Friday the 6th and Friday the 13th in the same month. The dotplot and computer output below summarize the data on the difference in the number of shoppers at each store on these two days (subtracting in the order 6th minus 13th).[25]



N	Mean	Median	TrMean	StDev	SEMean	Min	Max	Q1	Q3
45	-46.5	-11.0	-37.4	178.0	26.1	-774.0	302.0	-141.0	53.5

Researchers would like to carry out a test of $H_0: \mu_d = 0$ versus $H_a: \mu_d \neq 0$, where μ_d is the true mean difference in the number of grocery shoppers on these two days. Which of the following conditions for performing a paired t test are clearly satisfied?

I. Random **II.** 10% **III.** Normal/Large Sample

(a) I only (c) III only (e) I, II, and III

(b) II only (d) I and II only

100. The most important condition for sound conclusions from statistical inference is that

(a) the data come from a well-designed random sample or randomized experiment.

(b) the population distribution be exactly Normal.

(c) the data contain no outliers.

(d) the sample size be no more than 10% of the population size.

(e) the sample size be at least 30.

101. Vigorous exercise helps people live several years longer (on average). Whether mild activities like slow walking extend life is not clear. Suppose that the added life expectancy from regular slow walking is just 2 months. A statistical test is more likely to find a significant increase in mean life expectancy if

(a) it is based on a very large random sample and a 5% significance level is used.

(b) it is based on a very large random sample and a 1% significance level is used.

(c) it is based on a very small random sample and a 5% significance level is used.

(d) it is based on a very small random sample and a 1% significance level is used.

(e) the size of the sample doesn't have any effect on the significance of the test.

102. A researcher plans to conduct a significance test at the $\alpha = 0.01$ significance level. She designs her study to have a power of 0.90 at a particular alternative value of the parameter of interest. The probability that the researcher will commit a Type II error for the particular alternative value of the parameter at which she computed the power is

(a) 0.01. (b) 0.10. (c) 0.89. (d) 0.90. (e) 0.99.

103. Is your food safe? (8.1) "Do you feel confident or not confident that the food available at most grocery stores is safe to eat?" When a Gallup Poll asked this question, 87% of the sample said they were confident.[26] Gallup announced the poll's margin of error for 95% confidence as ±3 percentage points. Which of the following sources of error are included in this margin of error? Explain.

(a) Gallup dialed landline telephone numbers at random and so missed all people without landline phones, including people whose only phone is a cell phone.

(b) Some people whose numbers were chosen never answered the phone in several calls or answered but refused to participate in the poll.

(c) There is chance variation in the random selection of telephone numbers.

104. Spinning for apples (6.3 or 7.3) In the "Ask Marilyn" column of *Parade* magazine, a reader posed this question: "Say that a slot machine has five wheels, and each wheel has five symbols: an apple, a grape, a peach, a pear, and a plum. I pull the lever five times. What are the chances that I'll get at least one apple?" Suppose that the wheels spin independently and that the five symbols are equally likely to appear on each wheel in a given spin.

(a) Find the probability that the slot player gets at least one apple in one pull of the lever. Show your method clearly.

(b) Now answer the reader's question. Show your method clearly.

FRAPPY! Free Response AP® Problem, Yay!

The following problem is modeled after actual AP® Statistics exam free response questions. Your task is to generate a complete, concise response in 15 minutes.

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

Anne reads that the average price of regular gas in her state is $4.06 per gallon. To see if the average price of gas is different in her city, she selects 10 gas stations at random and records the price per gallon for regular gas at each station. The data, along with the sample mean and standard deviation, are listed in the table below.

Station	Price
1	4.13
2	4.01
3	4.09
4	4.05
5	3.97
6	3.99
7	4.05
8	3.98
9	4.09
10	4.02
Mean	**4.038**
SD	**0.0533**

Do the data provide convincing evidence that the average price of gas in Anne's city is different from $4.06 per gallon?

After you finish, you can view two example solutions on the book's Web site (www.whfreeman.com/tps5e). Determine whether you think each solution is "complete," "substantial," "developing," or "minimal." If the solution is not complete, what improvements would you suggest to the student who wrote it? Finally, your teacher will provide you with a scoring rubric. Score your response and note what, if anything, you would do differently to improve your own score.

Chapter Review

Section 9.1: Significance Tests: The Basics

In this section, you learned the basic ideas of significance testing. Start by stating the hypotheses that you want to test. The null hypothesis (H_0) is typically a statement of "no difference" and the alternative hypothesis (H_a) describes what we suspect is true. Remember that hypotheses are always about parameters, not statistics.

When sample data provide support for the alternative hypothesis, there are two possible explanations: (1) the null hypothesis is true, and data supporting the alternative hypothesis occurred just by chance, or (2) the alternative hypothesis is true, and the data are consistent with an alternative value of the parameter. In a significance test, always start with the belief that the null hypothesis is true. If you can rule out chance as a plausible explanation for the observed data, there is convincing evidence that the alternative hypothesis is true.

The *P*-value in a significance test measures how likely it is to get results at least as extreme as the observed results by chance alone, assuming the null hypothesis is true. To determine if the *P*-value is small enough to reject H_0, compare it to a predetermined significance level such as $\alpha = 0.05$. If *P*-value $< \alpha$, reject H_0—there is convincing evidence that the alternative hypothesis is true. However, if *P*-value $\geq \alpha$, fail to reject H_0—there is not convincing evidence that the alternative hypothesis is true.

Because conclusions are based on sample data, there is a possibility that the conclusion will be incorrect. You can make two types of errors in a significance test: a Type I error occurs if you find convincing evidence for the alternative hypothesis when, in reality, the null hypothesis is true. A Type II error occurs when you don't find convincing evidence that the alternative hypothesis is true when, in reality, the alternative hypothesis is true. The probability of making a Type I error is equal to the significance level (α) of the test.

Section 9.2: Tests about a Population Proportion

In this section, you learned the details of conducting a significance test for a population proportion p. Whenever you are asked if there is convincing evidence for a claim about a population proportion, you are expected to respond using the familiar four-step process.

STATE: Give the hypotheses you are testing in terms of p, state the significance level, and define the parameter p.

PLAN: Name the procedure you plan to use (one-sample z test for a population proportion) and check the appropriate conditions (Random, 10%, Large Counts) to see if the procedure is appropriate.

- Random: The data come from a well-designed random sample or randomized experiment.
  - **10%:** The sample size should be no larger than 10% of the population when sampling without replacement.
- Large Counts: Both np_0 and $n(1-p_0)$ must be at least 10, where p_0 is the value of p in the null hypothesis.

DO: Calculate the test statistic and *P*-value. The test statistic z measures how far away the sample statistic is from the hypothesized parameter value in standardized units:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

To calculate the *P*-value, use Table A or technology.

CONCLUDE: Use the *P*-value to make an appropriate conclusion about the hypotheses in context.

Perform a two-sided test when looking for convincing evidence that the true value of the parameter is *different* from the hypothesized value, in either direction. The *P*-value for a two-sided test is calculated by finding the probability of getting a sample statistic at least as extreme as the observed statistic, in either direction, assuming the null hypothesis is true.

You can also use a confidence interval to make a conclusion for a two-sided test. If the null parameter value is one of the plausible values in the interval, there isn't convincing evidence that the alternative hypothesis is true. However, if the null parameter value is not one of the plausible values in the interval, there is convincing evidence that the alternative hypothesis is true. Besides helping you draw a conclusion, the interval tells you which alternative parameter values are plausible.

The probability that you avoid making a Type II error when an alternative value of the parameter is true is called the power of the test. Power is good—if the alternative hypothesis is true, we want to maximize the probability of finding convincing evidence that it is true. We can increase the power of a significance test by increasing the sample size or by increasing the significance level. The power of a test will also be greater when the alternative value of the parameter is farther away from the null hypothesis value.

Section 9.3: Tests about a Population Mean

In this section, you learned the details of conducting a significance test for a population mean. Although some of the details are different, the reasoning and structure of the tests in this section are the same as in Section 9.2. In fact, the "State" and "Conclude" steps are exactly the same, other than the switch from proportions to means.

PLAN: Name the procedure you are using (one-sample t test for a population mean), and check the conditions (Random, 10%, and Normal/Large Sample). The Random and 10% conditions are the same as in Section 9.2. The Normal/Large Sample condition states that the population distribution must be Normal or the sample size must be large ($n \geq 30$). If the sample is small and the population shape is unknown, graph the sample data to make sure there is no strong skewness or outliers.

DO: Calculate the test statistic and P-value. The test statistic t measures how far away the sample statistic is from the hypothesized parameter value in standardized units:

$$t = \frac{\bar{x} - \mu}{s_x/\sqrt{n}}$$

To calculate the P-value, determine the degrees of freedom ($df = n - 1$) and use Table B or technology.

Use a paired t test to analyze the results of comparative experiments and observational studies that produce paired data. Start by calculating the difference for each pair and use the set of differences to check the Normal/Large Sample condition and to calculate the test statistic and P-value.

Remember to use significance tests wisely. When planning a study, use a large enough sample size so the test will have adequate power. Also, remember that statistically significant results aren't always "practically" important. Finally, be aware that the probability of making at least one Type I error goes up dramatically when conducting multiple tests.

What Did You Learn?

Learning Objective	Section	Related Example on Page(s)	Relevant Chapter Review Exercise(s)
State the null and alternative hypotheses for a significance test about a population parameter.	9.1	540	R9.1
Interpret a P-value in context.	9.1	543, 544	R9.5
Determine if the results of a study are statistically significant and draw an appropriate conclusion using a significance level.	9.1	546	R9.5
Interpret a Type I and a Type II error in context, and give a consequence of each.	9.1	548	R9.3, R9.4
State and check the Random, 10%, and Large Counts conditions for performing a significance test about a population proportion.	9.2	555	R9.4
Perform a significance test about a population proportion.	9.2	559, 562	R9.4
Interpret the power of a test and describe what factors affect the power of a test.	9.2	565, Discussion on 568	R9.3
Describe the relationship among the probability of a Type I error (significance level), the probability of a Type II error, and the power of a test.	9.2	565	R9.3
State and check the Random, 10%, and Normal/Large Sample conditions for performing a significance test about a population mean.	9.3	575	R9.2, R9.6, R9.7
Perform a significance test about a population mean.	9.3	580, 583	R9.6
Use a confidence interval to draw a conclusion for a two-sided test about a population parameter.	9.2, 9.3	563, 585	R9.5, R9.6
Perform a significance test about a mean difference using paired data.	9.3	586	R9.7

Chapter 9 Chapter Review Exercises

These exercises are designed to help you review the important ideas and methods of the chapter.

R9.1 Stating hypotheses State the appropriate null and alternative hypotheses in each of the following settings. Be sure to define the parameter.

(a) The average height of 18-year-old American women is 64.2 inches. You wonder whether the mean height of this year's female graduates from a large local high school (over 3000 students) differs from the national average. You measure an SRS of 48 female graduates and find that $\bar{x} = 63.1$ inches.

(b) Mr. Starnes believes that less than 75% of the students at his school completed their math homework last night. The math teachers inspect the homework assignments from a random sample of students at the school to help Mr. Starnes test his claim.

R9.2 Fonts and reading ease Does the use of fancy type fonts slow down the reading of text on a computer screen? Adults can read four paragraphs of text in the common Times New Roman font in an average time of 22 seconds. Researchers asked a random sample of 24 adults to read this text in the ornate font named Gigi. Here are their times, in seconds:

23.2	21.2	28.9	27.7	29.1	27.3	16.1	22.6	25.6	34.2	23.9	26.8
20.5	34.3	21.4	32.6	26.2	34.1	31.5	24.6	23.0	28.6	24.4	28.1

State and check the conditions for performing a significance test using these data.

R9.3 Strong chairs? A company that manufactures classroom chairs for high school students claims that the mean breaking strength of the chairs that they make is 300 pounds. One of the chairs collapsed beneath a 220-pound student last week. You wonder whether the manufacturer is exaggerating the breaking strength of the chairs.

(a) State appropriate null and alternative hypotheses in this setting. Be sure to define your parameter.

(b) Describe a Type I error and a Type II error in this setting, and give the consequences of each.

(c) Would you recommend a significance level of 0.01, 0.05, or 0.10 for this test? Justify your choice.

(d) The power of the test to detect $\mu = 294$ using $\alpha = 0.05$ is 0.71. Interpret this value in context.

(e) Explain two ways that you could increase the power of the test from (d).

R9.4 Flu vaccine A drug company has developed a new vaccine for preventing the flu. The company claims that fewer than 5% of adults who use its vaccine will get the flu. To test the claim, researchers give the vaccine to a random sample of 1000 adults. Of these, 43 get the flu.

(a) Do these data provide convincing evidence to support the company's claim?

(b) Which kind of mistake—a Type I error or a Type II error—could you have made in (a)? Explain.

(c) From the company's point of view, would a Type I error or Type II error be more serious? Why?

R9.5 Roulette An American roulette wheel has 18 red slots among its 38 slots. To test if a particular roulette wheel is fair, you spin the wheel 50 times and the ball lands in a red slot 31 times. The resulting *P*-value is 0.0384.

(a) Interpret the *P*-value in context.

(b) Are the results statistically significant at the $\alpha = 0.05$ level? Explain. What conclusion would you make?

(c) The casino manager uses your data to produce a 99% confidence interval for p and gets (0.44, 0.80). He says that this interval provides convincing evidence that the wheel is fair. How do you respond?

R9.6 Radon detectors Radon is a colorless, odorless gas that is naturally released by rocks and soils and may concentrate in tightly closed houses. Because radon is slightly radioactive, there is some concern that it may be a health hazard. Radon detectors are sold to homeowners worried about this risk, but the detectors may be inaccurate. University researchers placed a random sample of 11 detectors in a chamber where they were exposed to 105 picocuries per liter of radon over 3 days. A graph of the radon readings from the 11 detectors shows no strong skewness or outliers. The mean reading is 104.82 and the standard deviation of the readings is 9.54.

(a) Is there convincing evidence at the 10% level that the mean reading differs from the true value 105?

(b) A 90% confidence interval for the true mean reading is (99.61, 110.03). Is this interval consistent with your conclusion from part (a)? Explain.

R9.7 Better barley Does drying barley seeds in a kiln increase the yield of barley? A famous experiment by William S. Gosset (who discovered the *t* distributions) investigated this question. Eleven pairs of adjacent plots were marked out in a large field. For each pair, regular barley seeds were planted in one plot and kiln-dried seeds were planted in the other. The following table displays the data on yield (lb/acre).[27]

Plot	Regular	Kiln
1	1903	2009
2	1935	1915
3	1910	2011
4	2496	2463
5	2108	2180
6	1961	1925
7	2060	2122
8	1444	1482
9	1612	1542
10	1316	1443
11	1511	1535

(a) How can the Random condition be satisfied in this study?

(b) Assuming that the Random condition has been met, do these data provide convincing evidence that drying barley seeds in a kiln increases the yield of barley, on average? Justify your answer.

Chapter 9 AP® Statistics Practice Test

Section I: Multiple Choice *Select the best answer for each question.*

T9.1 An opinion poll asks a random sample of adults whether they favor banning ownership of handguns by private citizens. A commentator believes that more than half of all adults favor such a ban. The null and alternative hypotheses you would use to test this claim are

(a) $H_0: \hat{p} = 0.5; H_a: \hat{p} > 0.5$

(b) $H_0: p = 0.5; H_a: p > 0.5$

(c) $H_0: p = 0.5; H_a: p < 0.5$

(d) $H_0: p = 0.5; H_a: p \neq 0.5$

(e) $H_0: p > 0.5; H_a: p = 0.5$

T9.2 You are thinking of conducting a one-sample t test about a population mean μ using a 0.05 significance level. Which of the following statements is correct?

(a) You should not carry out the test if the sample does not have a Normal distribution.

(b) You can safely carry out the test if there are no outliers, regardless of the sample size.

(c) You can carry out the test if a graph of the data shows no strong skewness, regardless of the sample size.

(d) You can carry out the test only if the population standard deviation is known.

(e) You can safely carry out the test if your sample size is at least 30.

T9.3 To determine the reliability of experts who interpret lie detector tests in criminal investigations, a random sample of 280 such cases was studied. The results were

	Suspect's True Status	
Examiner's Decision	**Innocent**	**Guilty**
"Innocent"	131	15
"Guilty"	9	125

If the hypotheses are H_0: suspect is innocent versus H_a: suspect is guilty, then we could estimate the probability that experts who interpret lie detector tests will make a Type II error as

(a) 15/280. (c) 15/140. (e) 15/146.

(b) 9/280. (d) 9/140.

T9.4 A significance test allows you to reject a null hypothesis H_0 in favor of an alternative H_a at the 5% significance level. What can you say about significance at the 1% level?

(a) H_0 can be rejected at the 1% significance level.

(b) There is insufficient evidence to reject H_0 at the 1% significance level.

(c) There is sufficient evidence to accept H_0 at the 1% significance level.

(d) H_a can be rejected at the 1% significance level.

(e) The answer can't be determined from the information given.

T9.5 A random sample of 100 likely voters in a small city produced 59 voters in favor of Candidate A. The observed value of the test statistic for testing the null hypothesis $H_0: p = 0.5$ versus the alternative hypothesis $H_a: p > 0.5$ is

(a) $z = \dfrac{0.59 - 0.5}{\sqrt{\dfrac{0.59(0.41)}{100}}}$

(b) $z = \dfrac{0.59 - 0.5}{\sqrt{\dfrac{0.5(0.5)}{100}}}$

(c) $z = \dfrac{0.5 - 0.59}{\sqrt{\dfrac{0.59(0.41)}{100}}}$

(d) $z = \dfrac{0.5 - 0.59}{\sqrt{\dfrac{0.5(0.5)}{100}}}$

(e) $t = \dfrac{0.59 - 0.5}{\sqrt{\dfrac{0.5(0.5)}{100}}}$

T9.6 A researcher claims to have found a drug that causes people to grow taller. The coach of the basketball team at Brandon University has expressed interest but demands evidence. Over 1000 Brandon students volunteer to participate in an experiment to test this new drug. Fifty of the volunteers are randomly selected, their heights are measured, and they are given the drug. Two weeks later, their heights are measured again. The power of the test to detect an average increase in height of 1 inch could be increased by

(a) using only volunteers from the basketball team in the experiment.

(b) using $\alpha = 0.01$ instead of $\alpha = 0.05$.

(c) using $\alpha = 0.05$ instead of $\alpha = 0.01$.

(d) giving the drug to 25 randomly selected students instead of 50.

(e) using a two-sided test instead of a one-sided test.

T9.7 A 95% confidence interval for a population mean μ is calculated to be (1.7, 3.5). Assume that the conditions for performing inference are met. What conclusion can we draw for a test of $H_0: \mu = 2$ versus $H_a: \mu \neq 2$ at the $\alpha = 0.05$ level based on the confidence interval?

(a) None. We cannot carry out the test without the original data.

(b) None. We cannot draw a conclusion at the $\alpha = 0.05$ level because this test corresponds to the 97.5% confidence interval.

(c) None. Confidence intervals and significance tests are unrelated procedures.

(d) We would reject H_0 at level $\alpha = 0.05$.

(e) We would fail to reject H_0 at level $\alpha = 0.05$.

T9.8 In a test of $H_0: p = 0.4$ against $H_a: p \neq 0.4$, a random sample of size 100 yields a test statistic of $z = 1.28$. The P-value of the test is approximately equal to

(a) 0.90. (c) 0.05. (e) 0.10.

(b) 0.40. (d) 0.20.

T9.9 An SRS of 100 postal employees found that the average time these employees had worked at the postal service was 7 years with standard deviation 2 years. Do these data provide convincing evidence that the mean time of employment μ for the population of postal employees has changed from the value of 7.5 that was true 20 years ago? To determine this, we test the hypotheses $H_0: \mu = 7.5$ versus $H_a: \mu \neq 7.5$ using a one-sample t test. What conclusion should we draw at the 5% significance level?

(a) There is convincing evidence that the mean time working with the postal service has changed.

(b) There is not convincing evidence that the mean time working with the postal service has changed.

(c) There is convincing evidence that the mean time working with the postal service is still 7.5 years.

(d) There is convincing evidence that the mean time working with the postal service is now 7 years.

(e) We cannot draw a conclusion at the 5% significance level. The sample size is too small.

T9.10 Are TV commercials louder than their surrounding programs? To find out, researchers collected data on 50 randomly selected commercials in a given week. With the television's volume at a fixed setting, they measured the maximum loudness of each commercial and the maximum loudness in the first 30 seconds of regular programming that followed. Assuming conditions for inference are met, the most appropriate method for answering the question of interest is

(a) a one-proportion z test.

(b) a one-proportion z interval.

(c) a paired t test.

(d) a paired t interval.

(e) None of these.

Section II: Free Response *Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.*

T9.11 A software company is trying to decide whether to produce an upgrade of one of its programs. Customers would have to pay $100 for the upgrade. For the upgrade to be profitable, the company needs to sell it to more than 20% of their customers. You contact a random sample of 60 customers and find that 16 would be willing to pay $100 for the upgrade.

(a) Do the sample data give good evidence that more than 20% of the company's customers are willing to purchase the upgrade? Carry out an appropriate test at the $\alpha = 0.05$ significance level.

(b) Which would be a more serious mistake in this setting—a Type I error or a Type II error? Justify your answer.

(c) Describe two ways to increase the power of the test in part (a).

T9.12 "I can't get through my day without coffee" is a common statement from many students. Assumed benefits include keeping students awake during lectures and making them more alert for exams and tests. Students in a statistics class designed an experiment to measure memory retention with and without drinking a cup of coffee one hour before a test. This experiment took place on two different days in the same week (Monday and Wednesday). Ten students were used. Each student received no coffee or one cup of coffee one hour before the test on a particular day. The test consisted of a series of words flashed on a screen, after which the student had to write down as many of the words as possible. On the other day, each student received a different amount of coffee (none or one cup).

(a) One of the researchers suggested that all the subjects in the experiment drink no coffee before Monday's test and one cup of coffee before Wednesday's test. Explain to the researcher why this is a bad idea *and* suggest a better method of deciding when each subject receives the two treatments.

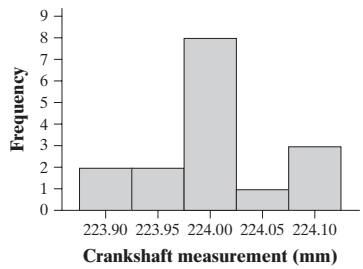
(b) The data from the experiment are provided in the table below. Set up and carry out an appropriate test to determine whether there is convincing evidence that drinking coffee improves memory.

Student	No cup	One cup
1	24	25
2	30	31
3	22	23
4	24	24
5	26	27
6	23	25
7	26	28
8	20	20
9	27	27
10	28	30

T9.13 A government report says that the average amount of money spent per U.S. household per week on food is about $158. A random sample of 50 households in a small city is selected, and their weekly spending on food is recorded. The sample data have a mean of $165 and a standard deviation of $20. Is there convincing evidence that the mean weekly spending on food in this city differs from the national figure of $158?

entered. **(b)** It is likely that more than 171 respondents have run red lights because some people may lie and say they haven't run a red light. The margin of error does not account for these sources of bias; it accounts only for sampling variability.

R8.7 **(a)** S: μ = the true mean measurement of the critical dimension for the engine crankshafts produced in one day. *P:* One-sample t interval for μ. Random: The data come from an SRS. 10%: the sample size (16) is less than 10% of all crankshafts produced in one day. Normal/Large Sample: the histogram shows no strong skewness or outliers.



D: Using df = 15, (223.969, 224.035). *C:* We are 95% confident that the interval from 223.969 to 224.035 mm captures the true mean measurement of the critical dimension for engine crankshafts produced on this day. **(b)** Because 224 is a plausible value in this interval, we don't have convincing evidence that the process mean has drifted.

R8.8 Solving $1.96\left(\frac{3000}{\sqrt{n}}\right) \le 1000$ gives $n \ge 35$.

R8.9 **(a)** The margin of error must get larger to increase the capture rate of the intervals. **(b)** If we quadruple the sample size, the margin of error will decrease by a factor of 2.

R8.10 **(a)** When we use the sample standard deviation s_x to estimate the population standard deviation σ. **(b)** The t distributions are wider than the standard Normal distribution and they have a slightly different shape with more area in the tails. **(c)** As the degrees of freedom increase, the spread and shape of the t distributions become more like the standard Normal distribution.

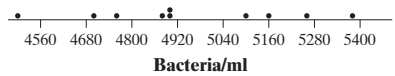
Answers to Chapter 8 AP® Statistics Practice Test

T8.1 a
T8.2 d
T8.3 c
T8.4 d
T8.5 b
T8.6 a
T8.7 c
T8.8 d
T8.9 e
T8.10 d

T8.11 **(a)** S: p = the true proportion of all visitors to Yellowstone who would say they favor the restrictions. *P:* One-sample z interval for p. Random: the visitors were selected randomly. 10%: the sample size (150) is less than 10% of all visitors to Yellowstone National Park. Large Counts: $n\hat{p} = 89 \ge 10$ and $n(1-\hat{p}) = 61 \ge 10$. *D:* (0.490, 0.696). *C:* We are 99% confident that the interval from 0.490 to 0.696 captures the true proportion of all visitors who would say that they favor the restrictions. **(b)** Because there are values smaller than 0.50 in the confidence interval, the U.S. Forest Service cannot conclude that more than half of visitors to Yellowstone National Park favor the proposal.

T8.12 **(a)** Because the sample size is large ($n = 48 \ge 30$), the Normal/Large Sample condition is met. **(b)** Maurice's interval uses a z critical value instead of a t critical value. Also, Maurice used the wrong value in the square root—it should be $n = 48$. *Correct:* Using df = 40, $6.208 \pm 2.021\left(\frac{2.576}{\sqrt{48}}\right) = (5.457, 6.959)$. *Using technology:* (5.46, 6.956) with df = 47.

T8.13 S: μ = the true mean number of bacteria per milliliter of raw milk received at the factory. *P:* One-sample t interval for μ. Random: The data come from a random sample. 10%: the sample size (10) is less than 10% of all 1-ml specimens that arrive at the factory. Normal/Large Sample: the dotplot shows that there is no strong skewness or outliers.



D: Using df = 9, (4794.37,5105.63). *C:* We are 90% confident that the interval from 4794.37 to 5105.63 bacterial/ml captures the true mean number of bacteria in the milk received at this factory.

Chapter 9

Section 9.1

Answers to Check Your Understanding

page 541: **1.** **(a)** p = proportion of all students at Jannie's high school who get less than 8 hours of sleep at night. **(b)** H_0:$p = 0.85$ and H_a:$p \ne 0.85$. **2.** **(a)** μ = true mean amount of time that it takes to complete the census form. **(b)** H_0:$\mu = 10$ and H_a:$\mu > 10$.

page 549: **1.** Finding convincing evidence that the new batteries last longer than 30 hours on average, when in reality their true mean lifetime is 30 hours. **2.** Not finding convincing evidence that the new batteries last longer than 30 hours on average, when in reality their true mean lifetime $>$30 hours. **3.** Answers will vary. A consequence of a Type I error would be that the company spends the extra money to produce these new batteries when they aren't any better than the older, cheaper type. A consequence of a Type II error would be that the company would not produce the new batteries, even though they were better.

Answers to Odd-Numbered Section 9.1 Exercises

9.1 H_0:$\mu = 115$; H_a:$\mu > 115$, where μ is the true mean score on the SSHA for all students at least 30 years of age at the teacher's college.

9.3 H_0:$p = 0.12$; H_a:$p \ne 0.12$, where p is the true proportion of lefties at his large community college.

9.5 H_0: $\sigma = 3$; H_a: $\sigma > 3$, where σ is the true standard deviation of the temperature in the cabin.

9.7 The null hypothesis is always that there is "no difference" or "no change" and the alternative hypothesis is what we suspect is true. Correct: H_0:$p = 0.37$; H_a:$p > 0.37$.

9.9 Hypotheses are always about population parameters. Correct: H_0:$\mu = 1000$ grams; H_a:$\mu < 1000$ grams.

9.11 **(a)** The attitudes of older students do not differ from other students, on average. **(b)** Assuming the mean score on the SSHA for students at least 30 years of age at this school is really 115, there is a 0.0101 probability of getting a sample mean of at least 125.7 just by chance in an SRS of 45 older students.

9.13 $\alpha = 0.10$: Because the P-value of $0.2184 > \alpha = 0.10$, we fail to reject H_0. We do not have convincing evidence that the proportion of left-handed students at Simon's college is different from the national proportion. $\alpha = 0.05$: Because the P-value of $0.2184 > \alpha = 0.05$, we fail to reject H_0. We do not have convincing evidence that the proportion of left-handed students at Simon's college is different from the national proportion.

9.15 $\alpha = 0.05$: Because the P-value of $0.0101 < \alpha = 0.05$, we reject H_0. We have convincing evidence that the true mean score on the SSHA for all students at least 30 years of age at the teacher's college > 115. $\alpha = 0.01$: Because the P-value of $0.0101 > \alpha = 0.01$, we fail to reject H_0. We do not have convincing evidence that the true mean score on the SSHA for all students at least 30 years of age at the teacher's college > 115.

9.17 Either H_0 is true or H_0 is false—it isn't true some of the time and not true at other times.

9.19 The P-value should be compared with a significance level (such as $\alpha = 0.05$), not the hypothesized value of p. Also, the data never "prove" that a hypothesis is true, no matter how large or small the P-value.

9.21 **(a)** H_0:$\mu = 6.7$; H_a:$\mu < 6.7$, where μ represents the mean response time for all accidents involving life-threatening injuries in the city. **(b)** I: Finding convincing evidence that the mean response time has decreased when it really hasn't. A consequence is that the city may not investigate other ways to reduce the mean response time and more people could die. II: Not finding convincing evidence that the mean response time has decreased when it really has. A consequence is that the city spends time and money investigating other methods to reduce the mean response time when they aren't necessary. **(c)** Type I, because people may end up dying as a result.

9.23 **(a)** H_0:$\mu = \$85{,}000$; H_a:$\mu > \$85{,}000$, where $\mu =$ the mean income of all residents near the restaurant. **(b)** I: Finding convincing evidence that the mean income of all residents near the restaurant exceeds $85,000 when in reality it does not. The consequence is that you will open your restaurant in a location where the residents will not be able to support it. II: Not finding convincing evidence that the mean income of all residents near the restaurant exceeds $85,000 when in reality it does. The consequence of this error is that you will not open your restaurant in a location where the residents would have been able to support it and you lose potential income.

9.25 d

9.27 c

9.29 **(a)** $P(\text{woman}) = 0.4168$, so $(24{,}611)(0.4168) = 10{,}258$ degrees were awarded to women. **(b)** No. $P(\text{woman}) = 0.4168$, which is not equal to $P(\text{woman} \mid \text{bachelors}) = 0.43$.

(c) $P(\text{at least 1 of the 2 degrees earned by a woman})$
$= 1 - P(\text{neither degree is earned by a woman}) =$

$$1 - \left(\frac{14{,}353}{24{,}611}\right)\left(\frac{14{,}352}{24{,}610}\right) = 0.6599$$

Section 9.2

Answers to Check Your Understanding

page 560: S: H_0:$p = 0.20$ versus H_a:$p > 0.20$, where p is the true proportion of all teens at the school who would say they have electronically sent or posted sexually suggestive images of themselves. P: One-sample z test for p. Random: Random sample. 10%: The sample size $(250) < 10\%$ of the 2800 students. Large Counts: $250(0.2) = 50 \geq 10$ and $250(0.8) = 200 \geq 10$. D: $z = \dfrac{0.252 - 0.20}{\sqrt{\dfrac{0.20(0.80)}{250}}} = 2.06$ and $P(Z \geq 2.06) = 0.0197$. C: Because the P-value of $0.0197 < \alpha = 0.05$, we reject H_0. We have convincing evidence that more than 20% of the teens in her school would say they have electronically sent or posted sexually suggestive images of themselves.

page 563: S: H_0:$p = 0.75$ versus H_a:$p \neq 0.75$, where p is the true proportion of all restaurant employees at this chain who would say that work stress has a negative impact on their personal lives. P: One-sample z test for p. Random: Random sample. 10%: The sample size $(100) < 10\%$ of all employees. Large Counts: $100(0.75) = 75 \geq 10$ and $100(0.25) = 25 \geq 10$.

D: $z = \dfrac{0.68 - 0.75}{\sqrt{\dfrac{0.75(0.25)}{100}}} = -1.62$ and $2P(Z \leq -1.62) = 0.1052$. C: Because the P-value of $0.1052 > \alpha = 0.05$, we fail to reject H_0. We do not have convincing evidence that the true proportion of all restaurant employees at this large restaurant chain who would say that work stress has a negative impact on their personal lives is different from 0.75.

page 564: The confidence interval given in the output includes 0.75, which means that 0.75 is a plausible value for the population proportion that we are seeking. So both the significance test (which didn't rule out 0.75 as the proportion) and the confidence interval give the same conclusion. The confidence interval, however, gives a range of plausible values for the population proportion instead of only making a decision about a single value.

page 569: **1.** A Type II error. If a Type I error occurred, they would reject a good shipment of potatoes and have to wait to get a new delivery. However, if a Type II error occurred, they would accept a bad batch and make potato chips with blemishes. This might upset consumers and decrease sales. To minimize the probability of a Type II error, choose a large significance level such as $\alpha = 0.10$

2. **(a)** Increase. Increasing α to 0.10 makes it easier to reject the null hypothesis, which increases power. **(b)** Decrease. Decreasing the sample size means we don't have as much information to use when making the decision, which makes it less likely to correctly reject H_0. **(c)** Decrease. It is harder to detect a difference of 0.02 $(0.10 - 0.08)$ than a difference of 0.03 $(0.11 - 0.08)$.

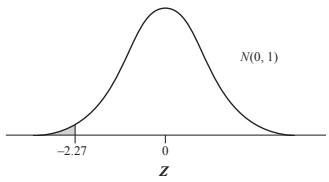
Answers to Odd-Numbered Section 9.2 Exercises

9.31 Random: Random sample. 10%: The sample size $(60) < 10\%$ of all students. Large Counts: $60\,(0.80) = 48 \geq 10$ and $60\,(0.20) = 12 \geq 10$.

9.33 $np_0 = 10(0.5) = 5$ and $n(1 - p_0) = 10(0.5) = 5$ are both < 10.

9.35 **(a)** $z = \dfrac{0.683 - 0.80}{\sqrt{\dfrac{0.80(0.20)}{60}}} = -2.27$ **(b)** $P(Z \leq -2.27) = 0.0116$.

Using technology: `normalcdf (lower:−1000, upper: −2.27, μ:0, σ:1) = 0.0116`. The graph is given below.



9.37 **(a)** P-value = 0.0143. 5%: Because the P-value of $0.0143 < \alpha = 0.05$, we reject H_0. There is convincing evidence that $p > 0.5$. 1%: Because the P-value of $0.0143 > \alpha = 0.01$, we fail to reject H_0. There is not convincing evidence that $p > 0.5$. **(b)** P-value = 0.0286. Because this P-value is still less than $\alpha = 0.05$ and greater than $\alpha = 0.01$, we would again reject H_0 at the 5% significance level and fail to reject H_0 at the 1% significance level.

9.39 *S:* H_0:$p = 0.37$ versus H_a:$p > 0.37$, where $p =$ true proportion of all students who are satisfied with the parking situation after the change. *P:* One-sample z test for p. Random: Random sample. 10%: The sample size (200) < 10% of the population of size 2500. Large Counts: $200(0.37) = 74 \geq 10$ and $200(0.63) = 126 \geq 10$. *D:* $z = 1.32$, P-value = 0.0934. *C:* Because the P-value of $0.0934 > \alpha = 0.05$, we fail to reject H_0. We do not have convincing evidence that the true proportion of all students who are satisfied with the parking situation after the change > 0.37.

9.41 **(a)** *S:* H_0:$p = 0.50$ versus H_a:$p > 0.50$, where p is the true proportion of boys among first-born children. *P:* One-sample z test for p. Random: Random sample. 10%: The sample size (25,468) < 10% of all first-borns. Large Counts: $25{,}468(0.50) = 12{,}734 \geq 10$ and $25{,}468(0.50) = 12{,}734 \geq 10$. *D:* $z = 5.49$, P-value≈0. *C:* Because the P-value of approximately $0 < \alpha = 0.05$, we reject H_0. There is convincing evidence that first-born children are more likely to be boys. **(b)** First-born children, because that is the group that we sampled from.

9.43 Here are the corrections: H_a: $p > 0.75$; $p =$ the true proportion of middle school students who engage in bullying behavior; 10%: the sample size (558) < 10% of the population of middle school students; $np_0 = 558(0.75) = 418.5 \geq 10$ and $n(1 - p_0) = 558(0.25) = 139.5 \geq 10$; $z = \dfrac{0.7975 - 0.75}{\sqrt{\dfrac{(0.75)(0.25)}{558}}} = 2.59$; P-value = 0.0048. Because the P-value of $0.0048 < \alpha = 0.05$, we reject H_0. We have convincing evidence that more than three-quarters of middle school students engage in bullying behavior.

9.45 *S:* H_0:$p = 0.60$ versus H_a:$p \neq 0.60$, where p is the true proportion of teens who pass their driving test on the first attempt. *P:* One-sample z test for p. Random: Random sample. 10%: The sample size (125) < 10% of all teens. Large Counts: $125(0.60) = 75 \geq 10$ and $125(0.40) = 50 \geq 10$. *D:* $z = 2.01$, P-value = 0.0444. *C:* Because our P-value of $0.0444 < \alpha = 0.05$, we reject H_0. There is convincing evidence that the true proportion of teens who pass the driving test on their first attempt is different from 0.60.

9.47 **(a)** *D:* (0.607,0.769). *C:* We are 95% confident that the interval from 0.607 to 0.769 captures the true proportion of teens who pass the driving test on the first attempt. **(b)** Because 0.60 is not in the interval, we have convincing evidence that the true proportion of teens who pass the driving test on their first attempt is different from 0.60.

9.49 No. Because the value 0.16 is included in the interval, we do not have convincing evidence that the true proportion of U.S. adults who would say they use Twitter differs from 0.16.

9.51 **(a)** $p =$ the true proportion of U.S. teens aged 13 to 17 who think that young people should wait to have sex until marriage. **(b)** Random: Random sample. 10%: The sample size (439) < 10% of the population of all U.S. teens. Large Counts: $439(0.5) = 219.5 \geq 10$ and $439(0.5) = 219.5 \geq 10$. **(c)** Assuming that the true proportion of U.S. teens aged 13 to 17 who think that young people should wait to have sex until marriage is 0.50, there is a 0.011 probability of getting a sample proportion that is at least as different from 0.5 as the proportion in the sample. **(d)** Yes. Because the P-value of $0.011 < \alpha = 0.05$, we reject H_0. There is convincing evidence that the true proportion of U.S. teens aged 13 to 17 who think that young people should wait to have sex until marriage differs from 0.5.

9.53 **(a)** I: Finding convincing evidence that more than 37% of students were satisfied with the new parking arrangement, when in reality only 37% were satisfied. *Consequence:* The principal believes that students are satisfied and takes no further action. II: Failing to find convincing evidence that more than 37% are satisfied with the new parking arrangement, when in reality more than 37% are satisfied. *Consequence:* The principal takes further action on parking when none is needed. **(b)** If the true proportion of students that are satisfied with the new arrangement is really 0.45, there is a 0.75 probability that the survey provides convincing evidence that the true proportion > 0.37. **(c)** Increase the sample size or significance level.

9.55 P(Type I) = $\alpha = 0.05$ and P(Type II) = 0.22.

9.57 **(a)** If the true proportion of Alzheimer's patients who would experience nausea is really 0.08, there is a 0.29 probability that the results of the study would provide convincing evidence that the true proportion < 0.10. **(b)** Increase the number of measurements taken (n) to get more information. **(c)** Decrease. If α is smaller, it becomes harder to reject the null hypothesis. This makes it harder to correctly reject H_0. **(d)** Increase. Because 0.07 is further from the null hypothesis value of 0.10, it will be easier to detect a difference between the null value and actual value.

9.59 c

9.61 b

9.63 **(a)** $X - Y$ has a Normal distribution with mean $\mu_{X-Y} = -0.2$ and standard deviation $\sigma_{X-Y} = \sqrt{(0.1)^2 + (0.05)^2} = 0.112$. To fit in a case, $X - Y$ must take on a negative number. **(b)** We want to find $P(X - Y < 0)$ using the $N(-0.2, 0.112)$ distribution. $z = \dfrac{0 - (-0.2)}{0.112} = 1.79$ and $P(Z < 1.79) = 0.9633$. *Using technology*: 0.9629. There is a 0.9629 probability that a randomly selected CD will fit in a randomly selected case. **(c)** P(all fit) = $(0.9629)^{100} = 0.0228$. There is a 0.0228 probability that all 100 CDs will fit in their cases.

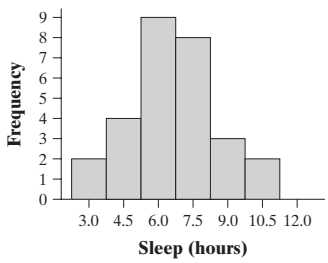
Section 9.3

Answers to Check Your Understanding

page 579: **1.** H_0:$\mu = 320$ versus H_a:$\mu \neq 320$, where $\mu =$ the true mean amount of active ingredient (in milligrams) in Aspro tablets from this batch of production. **2.** Random: Random sample. 10%: The sample of size 36 < 10% of the population of all tablets in this batch. Normal/Large Sample: $n = 36 \geq 30$. **3.** $t = \dfrac{319 - 320}{3/\sqrt{36}} = -2$ **4.** For this test, df = 35. Using Table B and df = 30, the tail area is between 0.025 and 0.05. Thus, the P-value for the two-sided test is between 0.05 and 0.10. *Using technology*: `2tcdf(lower:−1000,upper:−2,df:35) =`

$2(0.0267) = 0.0534$. Because the P-value of $0.0534 > \alpha = 0.05$, we fail to reject H_0. There is not convincing evidence that the true mean amount of the active ingredient in Aspro tablets from this batch of production differs from 320 mg.

page 583: **1.** S: $H_0{:}\mu = 8$ versus $H_a{:}\mu < 8$, where μ is the true mean amount of sleep that students at the professor's school get each night. P: One-sample t test for μ. Random: Random sample. 10%: The sample size (28) $<$ 10% of the population of students. Normal/Large Sample: The histogram below indicates that there is not much skewness and no outliers.



D: $\bar{x} = 6.643$ and $s_x = 1.981$. $t = -3.625$ and the P-value is between 0.0005 and 0.001. *Using technology*: P-value = 0.0006. C: Because our P-value of $0.0006 < \alpha = 0.05$, we reject H_0. There is convincing evidence that students at this university get less than 8 hours of sleep, on average.

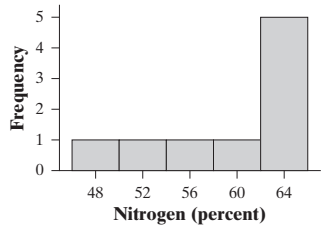
page 586: **1.** S: $H_0{:}\mu = 128$ versus $H_a{:}\mu \neq 128$, where μ is the true mean systolic blood pressure for the company's middle-aged male employees. P: One-sample t test for μ. Random: Random sample. 10%: The sample size (72) $<$ 10% of the population of middle-aged male employees. Normal/Large Sample: $n = 72 \geq 30$. D: $t = 1.10$ and P-value = 0.275. C: Because our P-value of $0.275 > \alpha = 0.05$, we fail to reject H_0. There is not convincing evidence that the mean systolic blood pressure for this company's middle-aged male employees differs from the national average of 128. **2.** We are 95% confident that the interval from 126.43 to 133.43 captures the true mean systolic blood pressure for the company's middle-aged male employees. The value of 128 is in this interval and therefore is a plausible mean systolic blood pressure for the males 35 to 44 years of age.

page 589: S: $H_0{:}\mu_d = 0$ versus $H_a{:}\mu_d > 0$, where μ_d is the true mean difference (air – nitrogen) in pressure lost. P: Paired t test for μ_d. Random: Treatments were assigned at random to each pair of tires. Normal/Large Sample: $n = 31 \geq 30$. D: $\bar{x} = 1.252$ and $s_x = 1.202$. $t = 5.80$ and P-value ≈ 0. C: Because the P-value of approximately $0 < \alpha = 0.05$, we reject H_0. We have convincing evidence that the true mean difference in pressure (air – nitrogen) > 0. In other words, we have convincing evidence that tires lose less pressure when filled with nitrogen than when filled with air, on average.

Answers to Odd-Numbered Section 9.3 Exercises

9.65 Random: Random sample. 10%: The sample size (45) $<$ 10% of the population size of 1000. Normal/Large Sample: $n = 45 \geq 30$.

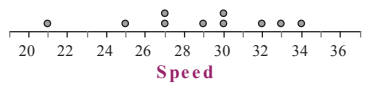
9.67 The Random condition may not be met, because we don't know if this is a random sample of the atmosphere in the Cretaceous era. Also, the Normal/Large Sample condition is not met. The sample size < 30 and the histogram below shows that the data are strongly skewed to the left.



9.69 **(a)** $t = \dfrac{125.7 - 115}{29.8/\sqrt{45}} = 2.409$. **(b)** For this test, df = 44. Using Table B and df = 40, we have $0.01 < P\text{-value} < 0.02$. *Using technology*: `tcdf(lower:2.409, upper:1000, df:44)` $= 0.0101$.

9.71 **(a)** Using Table B and df = 19, we have $0.025 < P\text{-value} < 0.05$. *Using technology*: P-value = 0.043. 5%: Because the P-value of $0.043 < \alpha = 0.05$, we reject H_0. There is convincing evidence that $\mu < 5$. 1%: Because the P-value of $0.043 > \alpha = 0.01$, we fail to reject H_0. There is not convincing evidence that $\mu < 5$. **(b)** *Using technology*: P-value = 0.086. 5%: Because the P-value of $0.086 > \alpha = 0.05$, we fail to reject H_0. There is not convincing evidence that $\mu \neq 5$. *1%:* same as part (a).

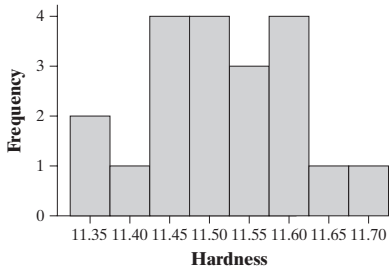
9.73 **(a)** S: $H_0{:}\ \mu = 25$ versus $H_a{:}\ \mu > 25$, where μ is the true mean speed of all drivers in a construction zone. P: One-sample t test for μ. Random: Random sample. 10%: The sample size (10) $<$ 10% of all drivers. Normal/Large Sample: There is no strong skewness or outliers in the sample.



D: $\bar{x} = 28.8$ and $s_x = 3.94$. $t = 3.05$, df = 9, and the P-value is between 0.005 and 0.01 (0.0069). C: Because the P-value of $0.0069 < \alpha = 0.05$, we reject H_0. We have convincing evidence that the true mean speed of all drivers in the construction zone > 25 mph. **(b)** Because we rejected H_0, it is possible we made a Type I error—finding convincing evidence that the true mean speed > 25 mph when it really isn't.

9.75 **(a)** S: $H_0{:}\mu = 1200$ versus $H_a{:}\mu < 1200$, where μ is the true mean daily calcium intake of women 18 to 24 years of age. P: One-sample t test for μ. Random: Random sample. 10%: The sample size (36) $<$ 10% of all women aged 18 to 24. Normal/Large Sample: $n = 36 \geq 30$. D: $t = -6.73$ and P-value = 0.000. C: Because the P-value of approximately $0 < \alpha = 0.05$, we reject H_0. There is convincing evidence that women aged 18 to 24 are getting less than 1200 mg of calcium daily, on average. **(b)** Assuming that women aged 18 to 24 get 1200 mg of calcium per day, on average, there is about a 0 probability that we would observe a sample mean ≤ 856.2 mg by chance alone.

9.77 S: $H_0{:}\mu = 11.5$ versus $H_a{:}\mu \neq 11.5$, where μ is the true mean hardness of the tablets. P: One-sample t test for μ. Random: The tablets were selected randomly. 10%: The sample size (20) $<$ 10% of all tablets in the batch. Normal/Large Sample: There is no strong skewness or outliers in the sample.



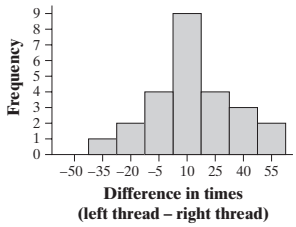
D: $\bar{x} = 11.5164$ and $s_x = 0.0950$. $t = 0.77$, df = 19, and the *P*-value is between 0.40 and 0.50 (0.4494). *C:* Because our *P*-value of $0.4494 > \alpha = 0.05$, we fail to reject H_0. We do not have convincing evidence that the true mean hardness of these tablets is different from 11.5.

9.79 *D:* With df = 19, (11.472,11.561). *C:* We are 95% confident that the interval from 11.472 to 11.561 captures the true mean hardness measurement for this type of pill. The confidence interval gives 11.5 as a plausible value for the true mean hardness μ, but it gives other plausible values as well.

9.81 *S:* H_0:$\mu = 200$ versus H_a:$\mu \neq 200$, where μ is the true mean response time of European servers. *P:* One-sample *t* interval to help us perform a two-sided test for μ. Random: The servers were selected randomly. 10%: The sample size (14) < 10% of all servers in Europe. Normal/Large Sample: The sample size is small, but a graph of the data reveals no strong skewness or outliers. *D:* (158.22, 189.64). *C:* Because our 95% confidence interval does not contain 200 milliseconds, we reject H_0 at the $\alpha = 0.05$ significance level. We have convincing evidence that the mean response time of European servers is different from 200 milliseconds.

9.83 **(a)** Yes. Because the *P*-value of $0.06 > \alpha = 0.05$, we fail to reject H_0: $\mu = 10$ at the 5% level of significance. Thus, the 95% confidence interval will include 10. **(b)** No. Because the *P*-value of $0.06 < \alpha = 0.10$, we reject H_0:$\mu = 10$ at the 10% level of significance. Thus, the 90% confidence interval would not include 10 as a plausible value.

9.85 **(a)** If all the subjects used the right thread first and they were tired when they used the left thread, then we wouldn't know if the difference in times was because of tiredness or because of the direction of the thread. **(b)** *S:* H_0: $\mu_d = 0$ versus H_a: $\mu_d > 0$, where μ_d is the true mean difference (left − right) in the time (in seconds) it takes to turn the knob with the left-hand thread and the right-hand thread. *P:* Paired *t* test for μ_d. Random: The order of treatments was determined at random. Normal/Large Sample: There is no strong skewness or outliers.



D: $\bar{x} = 13.32$ and $s_x = 22.94$. $t = 2.903$, df = 24, and the *P*-value is between 0.0025 and 0.005 (0.0039). *C:* Because the *P*-value of $0.0039 < \alpha = 0.05$, we reject H_0. We have convincing evidence that the true mean difference (left − right) in time it takes to turn the knob >0.

9.87 **(a)** H_0:$\mu_d = 0$ versus H_a:$\mu_d > 0$, where μ_d is the true mean difference in tomato yield (A − B). **(b)** df = 9. **(c)** *Interpretation*: Assuming that the average yield for both varieties is the same, there is a 0.1138 probability of getting a mean difference as large or larger than the one observed in this experiment. *Conclusion*: Because the *P*-value of $0.1138 > \alpha = 0.05$, we fail to reject H_0. We do not have convincing evidence that the true mean difference in tomato yield (A − B) > 0. **(d)** I: Finding convincing evidence that Variety A tomato plants have a greater mean yield, when in reality there is no difference. II: Not finding convincing evidence that Variety A tomato plants have a higher mean yield, when in reality Variety A does have a greater mean yield. They might have made a Type II error.

9.89 Increase the significance level α or increase the sample size *n*.

9.91 When the sample size is very large, rejecting the null hypothesis is very likely, even if the actual parameter is only slightly different from the hypothesized value.

9.93 **(a)** No, in a sample of size $n = 500$, we expect to see about $(500)(0.01) = 5$ people who do better than random guessing, with a significance level of 0.01. **(b)** The researcher should repeat the procedure on these four to see if they again perform well.

9.95 b

9.97 d

9.99 c

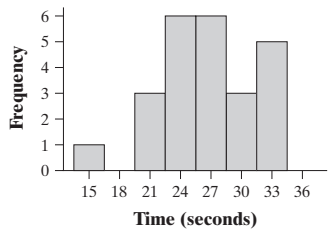
9.101 a

9.103 **(a)** Not included. The margin of error does not account for undercoverage. **(b)** Not included. The margin of error does not account for nonresponse. **(c)** Included. The margin of error is calculated to account for sampling variability.

Answers to Chapter 9 Review Exercises

R9.1 **(a)** H_0: $\mu = 64.2$; H_a: $\mu \neq 64.2$, where μ = the true mean height of this year's female graduates from the local high school. **(b)** H_0: $p = 0.75$; H_a: $p < 0.75$, where $p =$ the true proportion of all students at Mr. Starnes's school who completed their math homework last night.

R9.2 Random: Random sample. 10%: The sample size (24) < 10% of the population of adults. Normal/Large Sample: The histogram below shows that the distribution is roughly symmetric with no outliers.



R9.3 **(a)** H_0 : $\mu = 300$ versus H_a: $\mu < 300$, where μ = the true mean breaking strength of these chairs. **(b)** I: Finding convincing evidence that the mean breaking strength <300 pounds, when in reality it is 300 pounds or higher. *Consequence:* falsely accusing the company of lying. II: Not finding convincing evidence that the mean breaking strength <300 pounds, when in reality it <300 pounds. *Consequence:* allowing the company to continue to sell chairs that don't work as well as advertised. **(c)** Because a Type II error is more serious, increase the probability of a Type I error by using $\alpha = 0.10$. **(d)** If the true mean breaking strength is 294 pounds, there is a 0.71 probability that we will find convincing

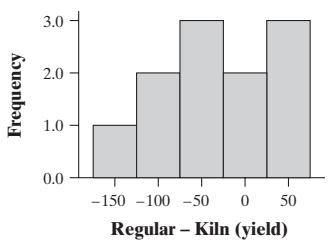
evidence that the true mean breaking strength < 300 pounds. **(e)** Increase the sample size or increase the significance level.

R9.4 **(a)** *S*: H_0: $p = 0.05$ versus H_a: $p < 0.05$, where p is the true proportion of adults who will get the flu after using the vaccine. *P*: One-sample z test for p. Random: Random sample. 10%: The sample size (1000) $< 10\%$ of the population of adults. Large Counts: $1000(0.05) = 50 \geq 10$ and $1000(0.95) = 950 \geq 10$. *D*: $z = -1.02$ and *P*-value $= 0.1539$. *C*: Because the *P*-value of $0.1539 > \alpha = 0.05$, we fail to reject H_0. We do not have convincing evidence that fewer than 5% of adults who receive this vaccine will get the flu. **(b)** Because we failed to reject the null hypothesis, we could have made a Type II error—not finding convincing evidence that the true proportion of adults get the flu after using this vaccine < 0.05, when in reality the true proportion < 0.05. **(c)** Answers will vary.

R9.5 **(a)** Assuming that the roulette wheel is fair, there is a 0.0384 probability that we would get a sample proportion of reds at least this different from the expected proportion of reds (18/38) by chance alone. **(b)** Because the *P*-value of $0.0384 < \alpha = 0.05$, the results are statistically significant at the $\alpha = 0.05$ level. This means that we reject H_0 and have convincing evidence that the true proportion of reds is different than $p = 18/38$. **(c)** Because $18/38 = 0.474$ is one of the plausible values in the interval, this interval does not provide convincing evidence that the wheel is unfair. It does not, however, prove that the wheel is fair as there are many other plausible values in the interval that are not equal to 18/38. Also, the conclusion here is inconsistent with the conclusion in part (b) because the manager used a 99% confidence interval, which is equivalent to a test using $\alpha = 0.01$.

R9.6 **(a)** *S*: H_0:$\mu = 105$ versus H_a:$\mu \neq 105$, where μ is the true mean reading from radon detectors. *P*: One-sample t test for μ. Random: Random sample. 10%: The sample size (11) $< 10\%$ of all radon detectors. Normal/Large Sample: A graph of the data shows no strong skewness or outliers. *D*: $t = -0.06$, df $= 10$, and and *P*-value > 0.50 (0.9513). *C*: Because the *P*-value of $0.9513 > \alpha = 0.10$, we fail to reject H_0. We do not have convincing evidence that the true mean reading from the radon detectors is different than 105. **(b)** Yes. Because 105 is in the interval from 99.61 to 110.03, both the confidence interval and the significance test agree that 105 is a plausible value for the true mean reading from the radon detectors.

R9.7 **(a)** The random condition can be satisfied by randomly allocating which plot got the regular barley seeds and which one got the kiln-dried seeds within each pair of adjacent plots. **(b)** *S*: H_0: $\mu_d = 0$ versus H_a: $\mu_d < 0$, where μ_d is the true mean difference (regular − kiln) in yield between regular barley seeds and kiln-dried barley seeds. *P*: Paired t test for μ_d. Random: Assumed. Normal/Large Sample: The histogram below shows no strong skewness or outliers.



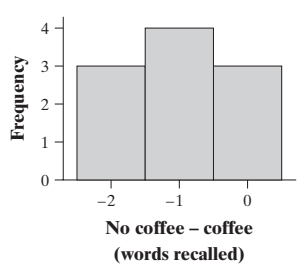
D: $\bar{x} = -33.7$ and $s_x = 66.2$. $t = -1.690$, df $= 10$, and the *P*-value is between 0.05 and 0.10 (0.0609). *C*: Because the *P*-value of $0.0609 > \alpha = 0.05$, we fail to reject H_0. We do not have convincing evidence that the true mean difference (regular − kiln) in yield < 0.

Answers to Chapter 9 AP® Statistics Practice Test

T9.1 b
T9.2 e
T9.3 c
T9.4 e
T9.5 b
T9.6 c
T9.7 e
T9.8 d
T9.9 a
T9.10 c

T9.11 **(a)** *S*: H_0: $p = 0.20$ versus H_a: $p > 0.20$, where p is the true proportion of customers who would pay \$100 for the upgrade. *P*: One-sample z test for p. Random: Random sample. 10%: The sample size (60) $< 10\%$ of this company's customers. Large Counts: $60(0.20) = 12 \geq 10$ and $60(0.8) = 48 \geq 10$. *D*: $z = 1.29$, *P*-value $= 0.0984$. *C*: Because the *P*-value of $0.0984 > \alpha = 0.05$, we fail to reject H_0. We do not have convincing evidence that more than 20% of customers would pay \$100 for the upgrade. **(b)** I: Finding convincing evidence that more than 20% of customers would pay for the upgrade, when in reality they would not. II: Not finding convincing evidence that more than 20% of customers would pay for the upgrade, when in reality more than 20% would. For the company, a Type I error is worse because they would go ahead with the upgrade and lose money. **(c)** Increase the sample size or increase the significance level.

T9.12 **(a)** Students may improve from Monday to Wednesday just because they have already done the task once. Then we wouldn't know if the experience with the test or the caffeine is the cause of the difference in scores. A better way to run the experiment would be to randomly assign half the students to get 1 cup of coffee on Monday and the other half to get no coffee on Monday. Then have each person do the opposite treatment on Wednesday. **(b)** *S*: H_0: $\mu_d = 0$ versus H_a: $\mu_d < 0$, where μ_d is the true mean difference (no coffee − coffee) in the number of words recalled without coffee and with coffee. *P*: Paired t test for μ_d. Random: The treatments were assigned at random. Normal/Large Sample: The histogram below shows a symmetric distribution with no outliers.



D: $\bar{x} = -1$ and $s_x = 0.816$. $t = -3.873$, df $= 9$, and the *P*-value is between 0.001 and 0.0025 (0.0019). *C*: Because the *P*-value of $0.0019 < \alpha = 0.05$, we reject H_0. We have convincing evidence that the mean difference (no coffee − coffee) in word recall < 0.

T9.13 *S*: H_0: $\mu = \$158$ versus H_a: $\mu \neq \$158$, where μ is the true mean amount spent on food by households in this city. *P*: One-sample t test for μ. Random: Random sample. 10%: The sample size (50) $< 10\%$ of households in this small city. Normal/Large

Sample: $n = 50 \geq 30$. *D*: $t = 2.47$; using df = 40, the *P*-value is between 0.01 and 0.02 (using df = 49, 0.0168). *C*: Because the *P*-value of $0.0168 < \alpha = 0.05$, we reject H_0. We have convincing evidence that the true mean amount spent on food per household in this city is different from the national average of $158.

Chapter 10

Section 10.1

Answers to Check Your Understanding

page 619: *S*: p_1 = true proportion of teens who go online every day and p_2 = true proportion of adults who go online every day. *P*: Two-sample *z* interval for $p_1 - p_2$. Random: Independent random samples. 10%: $n_1 = 799 < 10\%$ of teens and $n_2 = 2253 < 10\%$ of adults. Large Counts: 503, 296, 1532, and 721 are all ≥ 10. *D*: $(0.63 - 0.68) \pm 1.645\sqrt{\frac{0.63(0.37)}{799} + \frac{0.68(0.32)}{2253}} =$ $(-0.0824, -0.0176)$. *C*: We are 90% confident that the interval from -0.0824 to -0.0176 captures the true difference in the proportion of U.S. adults and teens who go online every day.

page 628: *S*: H_0: $p_1 - p_2 = 0$ versus H_a: $p_1 - p_2 > 0$, where p_1 is the true proportion of children like the ones in the study who do not attend preschool that use social services later and p_2 is the true proportion of children like the ones in the study who attend preschool that use social services later. *P*: Two-sample *z* test for $p_1 - p_2$. Random: Two groups in a randomized experiment. Large Counts: 49, 12, 38, 24 are all ≥ 10. *D*: $z = \frac{(0.8033 - 0.6129) - 0}{\sqrt{\frac{0.7073(0.2927)}{61} + \frac{0.7073(0.2927)}{62}}} = 2.32$ and *P*-value = 0.0102. *C*: Because the *P*-value of $0.0102 < \alpha = 0.05$, we reject H_0. There is convincing evidence that the true proportion of children like the ones in the study who do not attend preschool that use social services later is greater than the true proportion of children like the ones in the study who attend preschool that use social services later.

Answers to Odd-Numbered Section 10.1 Exercises

10.1 **(a)** Approximately Normal because $100(0.25) = 25$, $100(0.75) = 75$, $100(0.35) = 35$, and $100(0.65) = 65$ are all at least 10. **(b)** $\mu_{\hat{p}_1 - \hat{p}_2} = 0.25 - 0.35 = -0.10$. **(c)** Because $n_1 = 100 < 10\%$ of the first bag and $n_2 = 100 < 10\%$ of the second bag, $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{0.25(0.75)}{100} + \frac{0.35(0.65)}{100}} = 0.0644$.

10.3 **(a)** Approximately Normal because $50(0.30) = 15$, $50(0.7) = 35$, $100(0.15) = 15$, and $100(0.85) = 85$ are all at least 10. **(b)** $\mu_{\hat{p}_C - \hat{p}_A} = 0.30 - 0.15 = 0.15$. **(c)** Because $n_C = 50 < 10\%$ of the jelly beans in the Child mix and $n_A = 100 < 10\%$ of the jelly beans in the Adult mix, $\sigma_{\hat{p}_C - \hat{p}_A} = \sqrt{\frac{0.3(0.7)}{50} + \frac{0.15(0.85)}{100}} = 0.0740$.

10.5 The data do not come from independent random samples or two groups in a randomized experiment. Also, there were less than 10 successes (3) in the group from the west side of Woburn.

10.7 There were less than 10 failures (0) in the treatment group, less than 10 successes (8) in the control group, and less than 10 failures in the control group (4).

10.9 **(a)** $\text{SE}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{0.26(1-0.26)}{316} + \frac{0.14(1-0.14)}{532}} = 0.0289$. If we were to take many random samples of 316 young adults and 532 older adults, the difference in the sample proportions of young adults and older adults who use Twitter will typically be 0.0289 from the true difference. **(b)** *S*: p_1 = true proportion of young adults who use Twitter and p_2 = true proportion of older adults who use Twitter. *P*: Two-sample *z* interval for $p_1 - p_2$. Random: Two independent random samples. 10%: $n_1 = 316 < 10\%$ of all young adults and $n_2 = 532 < 10\%$ of all older adults. Large Counts: 82, 234, 74, 458 are all at least 10. *D*: (0.072,0.168). *C*: We are 90% confident that the interval from 0.072 to 0.168 captures the true difference in the proportions of young adults and older adults who use Twitter.

10.11 **(a)** *S*: p_1 = true proportion of young men who live in their parents' home and p_2 = true proportion of young women who live in their parents' home. *P*: Two-sample *z* interval for $p_1 - p_2$. Random: Reasonable to consider these independent random samples. 10%: $n_1 = 2253 < 10\%$ of the population of young men and $n_2 = 2629 < 10\%$ of the population of young women. Large Counts: 986, 1267, 923, 1706 are all at least 10. *D*: (0.051,0.123). *C*: We are 99% confident that the interval from 0.051 to 0.123 captures the true difference in the proportions of young men and young women who live in their parents' home. **(b)** Because the interval does not contain 0, there is convincing evidence that the true proportion of young men who live in their parents' home is different from the true proportion of young women who live in their parents' home.

10.13 $H_0: p_1 - p_2 = 0$ versus $H_a: p_1 - p_2 \neq 0$, where p_1 is the true proportion of all teens who would say that they own an iPod or MP3 player and p_2 is the true proportion of all young adults who would say that they own an iPod or MP3 player.

10.15 *P*: Two-sample *z* test for $p_1 - p_2$. Random: Independent random samples. 10%: $n_1 = 800 < 10\%$ of all teens and $n_2 = 400 < 10\%$ of all young adults. Large Counts: 632, 168, 268, and 132 are all at least 10. $D: z = 4.53$ and *P*-value ≈ 0. *C*: Because the *P*-value of close to $0 < \alpha = 0.05$, we reject H_0. There is convincing evidence that the true proportion of teens who would say that they own an iPod or MP3 player is different from the true proportion of young adults who would say that they own an iPod or MP3 player.

10.17 *D* : (0.066,0.174). *C* : We are 95% confident that the interval from 0.066 to 0.174 captures the true difference in proportions of teens and young adults who own iPods or MP3 players. Because 0 is not included in the interval, it is consistent with the results of Exercise 15.

10.19 *S* : $H_0: p_1 - p_2 = 0$ versus $H_a: p_1 - p_2 > 0$, where p_1 is the true proportion of 6- to 7-year-olds who would sort correctly and p_2 is the true proportion of 4- to 5-year-olds who would sort correctly. *P*: Two-sample *z* test for $p_1 - p_2$. Random: Independent random samples. 10%: $n_1 = 53 < 10\%$ of all 6- to 7-year olds and $n_2 = 50 < 10\%$ of all 4- to 5-year-olds. Large Counts: 28, 25, 10, 40 are all ≥ 10. *D*: $z = 3.45$ and *P*-value = 0.0003. *C*: Because the *P*-value of $0.0003 < \alpha = 0.05$, we reject H_0. We have convincing evidence that the true proportion of 6- to 7-year-olds who would sort correctly is greater than the true proportion of 4- to 5-year-olds who would sort correctly.

10.21 **(a)** *S*: $H_0: p_A - p_B = 0$ versus $H_a: p_A - p_B > 0$, where p_A is the true proportion of students like these who would pass the driver's license exam when taught by instructor A and p_B is the true